MEDICAL BIOPHYSICS

FOR 1ST YEAR MEDICAL STUDENTS

Octavian Călinescu ◆ Ramona Babeş ◆ Adrian Iftime
Irina Băran ◆ Diana Ionescu ◆ Constanţa Ganea

coordonator: Octavian Călinescu

Presa Universitară Clujeană

MEDICAL BIOPHYSICS

For 1st Year Medical Students

•

Coordonator:

OCTAVIAN CĂLINESCU

AUTORI:

Octavian Călinescu ◆ Ramona Babeş ◆ Adrian Iftime

Irina Băran ◆ Diana Ionescu ◆ Constanța Ganea

AUTHORS

Octavian Călinescu, PhD
Associate Professor
Department of Biophysics
Faculty of Medicine
"Carol Davila" University of Medicine and Pharmacy Bucharest

Ramona Babeș, PhD
Lecturer
Department of Biophysics
Faculty of Medicine
"Carol Davila" University of Medicine and Pharmacy Bucharest

Adrian Iftime, MD, PhD
Associate Professor
Department of Biophysics
Faculty of Medicine
"Carol Davila" University of Medicine and Pharmacy Bucharest

Irina Băran, PhD
Professor
Department of Biophysics
Faculty of Medicine
"Carol Davila" University of Medicine and Pharmacy Bucharest

Diana Ionescu, PhD
Associate Professor
Department of Biophysics
Faculty of Medicine
"Carol Davila" University of Medicine and Pharmacy Bucharest

Constanța Ganea, PhD
Professor
Department of Biophysics
Faculty of Medicine
"Carol Davila" University of Medicine and Pharmacy Bucharest

SCIENTIFIC REFEREES

Maria-Magdalena Mocanu, PhD
Professor
Department of Biochemistry and Biophysics
Faculty of Midwifery and Nursing
"Carol Davila" University of Medicine and Pharmacy Bucharest

Violeta Ristoiu, PhD
Professor
Department of Anatomy, Animal Physiology and Biophysics
Faculty of Biology
University of Bucharest

MEDICAL BIOPHYSICS

For 1st Year Medical Students

Coordonator:

OCTAVIAN CĂLINESCU

AUTORI:

Octavian Călinescu ◆ Ramona Babeş ◆ Adrian Iftime
Irina Băran ◆ Diana Ionescu ◆ Constanța Ganea

PRESA UNIVERSITARĂ CLUJEANĂ

2024

Referenți științifici:

Prof. univ. dr. Maria-Magdalena Mocanu

Prof. univ. dr. Violeta Ristoiu

ISBN 978-606-37-2235-6

© 2024 Coordonatorul volumului. Toate drepturile rezervate. Reproducerea integrală sau parțială a textului, prin orice mijloace, fără acordul coordonatorului, este interzisă şi se pedepsește conform legii.

Universitatea Babeș-Bolyai
Presa Universitară Clujeană
Director: Codruța Săcelean
Str. Hasdeu nr. 51
400371 Cluj-Napoca, România
Tel./fax: (+40)-264-597.401
E-mail: editura@ubbcluj.ro
http://www.editura.ubbcluj.ro/

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION TO MEDICAL BIOPHYSICS

It is initially difficult for most medical students to understand why they have to study biophysics. After all, high school biology textbooks make very little mention of physics, not to mention its dreaded cousin, mathematics.

We have become used to our world being divided into clearly delimited sections: living things are covered by biology, acids being corrosive is the problem of chemistry, while an object falling to the ground is dealt with by the laws of physics. We make a point of training specialists in all these fields, but the reason is not because they are completely different, but because, at the moment, the volume of human knowledge has become so high that it is impossible for one human being to be familiar with all the information in a given field. However, reality is not so clear-cut.

Let's consider an action that you are doing right now and dissect it for a bit. You are now reading these words, either printed on a piece of paper, or shown on a screen. This apparently simple act of you seeing what is written here is an incredibly complex process.

We see this text because incredibly tiny flecks of light called photons are reflected from the paper (or emitted by the backlight of the screen) and travel towards our eyes. In our eye, this light is deflected (refracted we call it in physics) and then directed and focused on the back of the eye, in a region called the retina. In the retina, light interacts with a layer of light sensitive neurons called photoreceptor cells. These cells have an electrical current that flows through their membrane while they are not exposed to light: the "dark current". When photons reach a photoreceptor cell, they trigger a set of chemical reactions that starts with the isomerization of a small molecule called retinal and eventually ends with the flow of the dark current stopping, a change which is then communicated to other neurons to which the photoreceptor cell "talks to" by release of molecules of glutamate. This information is then directed, electrically, to the visual cortex in the brain. And all that so you can read a few words on a page...

I hope you realize now that this biological process of vision is terribly difficult to describe with the tools of biology alone: we are talking about a physical phenomenon (light) that elicits a chemical response (isomerization), which is then translated into a physical response (a change in an electrical current). And this is just one example of many that shows why a medical doctor can only fully understand living organisms if he has good knowledge of physics and chemistry.

Let's now address some of the more common questions that you will have to start with.

Who is this book indended for?

This book was written for first year medical students, in particular, those studying medicine in English at the "Carol Davila" University of Medicine and Pharmacy Bucharest. However, even if you, the current reader, are not in that position, this book might still be for you if are interested in biophysics or the functioning of the human body in general.

What is biophysics?

Biophysics is a branch of science that deals with physical processes happening in living organisms and with using the tools of physics in order to explain biological processes. It is an interdisciplinary (or "border") science (Figure 1.1) that takes elements of biology, physics and chemistry. Indeed, the authors of this book, by training, are a chemist, a biologist, a medical doctor and three physicists.

Try to remember something from now on – everything you will study about the human body

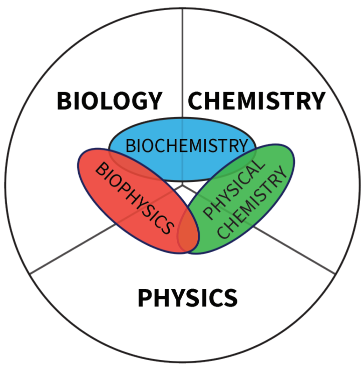


Figure 1.1. Biophysics is an interdisciplinary science.

in the different topics in medical school is connected. When you study biophysics this semester remember to cross-reference the information you get here with that you learn in biochemistry or physiology, for instance. If something seems like a contradiction, please let us know during the lectures – there has to be an error somewhere, because in the end we are all talking about the same thing – the living world.

What will I learn from this book?

We will present to you the physical basis of a large number of biological processes. Due to the limited number of lectures (only a semester), this book cannot be and is not an encyclopedia that covers all that there is to know about biophysics. However, we will present to you, just to list a few:

- how our cells transport molecules and ions across their membranes;
- how these cells use the movement of ions to impose an electrical potential difference across their membrane that is essential to their survival;
- how our muscles contract;
- how we hear and see;
- the effects of radiation on the human body;
- how we can visualize structures inside the human body without resorting to invasive surgery.

Do I need to know a lot of physics, mathematics, or chemistry beforehand?

The short answer is – yes, a bit, but not as much as you might fear. We will assume, though, that you know some notions before you read each chapter (these will be presented in the beginning of each chapter). For a refresher course in case you don't remember a particular concept, we recommend high school level textbooks. We will also recommend a few online resources at the end of this chapter.

Regarding mathematics, this is many times a "sensitive" topic for medical doctors. You will definitely need to use some mathematics to successfully study for biophysics, and, indeed, for your medical career. To put it bluntly – having mathematical abilities is essential for your future career. How can you calculate the dose of a drug that you want to administer or how can you interpret the results of a medical study without any knowledge of mathematics or statistics?

Mathematics is nothing to be afraid of, it's simply a tool that helps us interpret the physical reality around us. To quote Richard Feynman on the importance of mathematics in physics: "if you want to discuss nature, to learn about nature, and to appreciate nature, it's necessary to find out the language that she speaks in." That language, in case it was not obvious, is mathematics.

That being said, we will try to limit the needed mathematical knowledge in this text as much as possible. Very likely, if you have studied mathematics in high school for at least two years you will have no trouble with the level of mathematics in this book.

Should I just memorize everything in this book?

You could do that, and you'd probably also pass the exam. However, we encourage you to ask questions and understand the concepts presented in this book. Unlike some other topics you will study in medical school, little is gained by memorizing most of the equations in this book, for instance. Instead, try to figure out what that equation wants to present as a concept.

Note that only the parts of a chapter that are not marked as complementary information will be required at the exam (see below).

Complementary information

The topics presented in this book were summarized as best as possible. Sometimes, however, additional information is required in order to understand certain topics better. Also, there might be some among you which are curious to learn more.

In order to not burden you with this additional information for the exam, we chose to mark some sections as **complementary information** using a box with a particular shading, as you can see for this paragraph. All this information is still important, and sometimes critical for understanding the text. However, we will not ask it from you at the exam. Complementary information can be demonstrations, intricacies of a particular technique, detailed examples, extra figures, etc.

We strongly encourage you to read all the text (including the complementary information) at least once, ideally just before or just after each lecture. On further rereads, especially if you understood the explanations on the first go, you can skip the complementary information paragraphs or pages. You can return to them also in later years of study or when you encounter a particular issue during your medical practice.

So, what resources do you recommend in case I don't understand something in this book?

We will list here a few textbooks or online resources that will help you with some things that might not be presented in detail here. We would like to stress that you are not required to consult these resources, nor will you get any questions at

the exam from them. These merely serve as additional aid in case you want to understand more:

- A medical dictionary, for referencing medical terms you are not familiar with. Stedman's Medical Dictionary (stedmansonline.com) is highly recommended, but it is not free. Alternatively, a freely accessible resource is The Free Dictionary's Medical dictionary (medical-dictionary.thefreedictionary.com);
- Khan Academy (khanacademy.org) hosts numerous courses on science topics, including some of those discussed in this book;
- Light and Matter (lightandmatter.com) is a free textbook by Benjamin Cromwell which aims to give an introduction to general physics topics;
- Openstax hosts free textbooks for chemistry (openstax.org/details/books/chemistry-2e) and physics (openstax.org/details/books/physics);
- The Physics Classroom (physicsclassroom.com) hosts interactive simulations of some physical phenomena;
- Deranged Physiology (derangedphysiology.com) is an online resource by Alex Yartsev that tackles some issues in physiology that are also studied by biophysics;
- Hyperphysics presents basic and advanced topics in physics through the form of interconnected short segments (hyperphysics.phy-astr.gsu.edu);
- The Feynman Lectures on Physics collect Nobel laureate Richard Feynman's teaching of many principles of physics in an accessible manner (feynmanlectures.caltech.edu). If you have a curiosity for physics in general, we highly recommend this resource. However, take note that the level of physics presented there is much more advanced than that of this book.

That being said, we wish you luck in the study of Medical Biophysics! We hope you will enjoy it and find it useful later down the line in your career!

The authors

CHAPTER 2

BIOLOGICAL THERMODYNAMICS

Prerequisite knowledge

- Discrete structure of matter (atoms and molecules)
- Units of measurement, the International System of Units
- Energy
- Converting between different units of measurement, multiples, submultiples

1. GENERAL CONCEPTS OF THERMODYNAMICS

One way of looking at the human body is as an incredibly complex, integrated factory. With input of a few key ingredients (oxygen, water, nutrients), our bodies ensure the function of all the processes of life, including movement, growth, homeostasis or reproduction. A general law of the universe that will be described in this chapter is that the energy required for a certain action cannot come from nothing. For example, the movement of one's limbs requires the contraction of skeletal muscles, which is powered in the muscle fibers by the breakdown of energy-rich molecules of ATP. In turn, ATP is synthesized in our cells mainly by using the energy provided by another energy-rich molecule, glucose. We can thus say that chemical energy, initially stored in the form of glucose, then in ATP, is converted into mechanical energy (movement).

Let us first define **energy**, which in thermodynamics has a very simple definition: it is the ability to do work or produce heat. We can make this definition even simpler, energy acts like a "currency" that an object possesses. When the object needs to perform work or transfer heat, it needs to spend a part of its energy in order to do so.

The branch of physics that describes the ways energy is converted between its different forms is called **thermodynamics**. As previously stated, these transformations obey a set of physical laws, the laws (principles) of thermodynamics. Biological thermodynamics is a field that studies how living organisms cope with these physical laws, how energy is generated, transformed and spent in the body, and the energy-based relations of the organism with the surrounding environment. The tools used in physics for measuring energy transfers are also applied in biology.

This chapter aims to give you a brief overview of the main concepts of thermodynamics and to point out some applications of thermodynamics in medicine and living organisms in general. We will begin by defining a few notions that are indispensable for understanding the rest of this chapter.

1.1. Thermodynamic systems

Thermodynamics deals with assemblies which are made up of a large number of particles that interact with each other and with their environment and have well-defined limits. These are called **thermodynamic systems**. Virtually anything in the physical world is definable as a thermodynamic system, as long as it fulfils these conditions. For example, a coffee cup, a book, a room, a computer or a living being are all thermodynamic systems. On the other hand, an atom or a molecule are not thermodynamic systems, as they have too few particles.

As stated from their definition, thermodynamic systems interact with their environment. Depending on the types of interactions possible, we can classify (Figure 2.1) thermodynamic systems as follows:

- **Open systems** can exchange both substance and energy with their environment. An open bottle is a good example of such a system. Living organisms are open thermodynamic systems.
- **Closed systems** can exchange only energy, but not substance, with their environment. A closed bottle is an example of such a system.

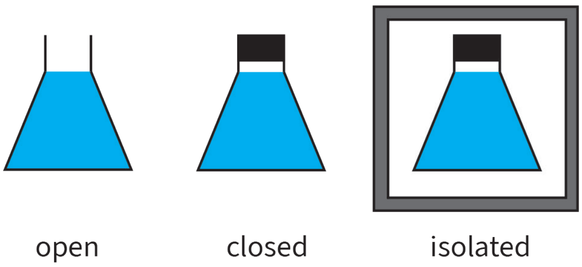


Figure 2.1. Classification of thermodynamic systems.

- **Isolated systems** do not exchange substance or energy with their environment. These have walls that prevent the transfer of heat to and from the medium (adiabatic walls) and are a convenient system to study, as they are simpler to describe. In reality, an ideally isolated system does not exist. An approximation of such a system is an insulated bottle (called also thermos bottle or Dewar flask) meant to keep liquids hot for extended periods of time – the insulated bottle has walls that slow down the loss of heat of the liquid, but cannot prevent it completely. Hence, your coffee will still get cold even in an insulated bottle, it just takes a little more time.

1.2. Thermodynamic parameters

In order to quantitatively describe a thermodynamic system *at a particular point in time*, one needs to be able to measure and define certain properties (physical quantities) of the system *at that particular time*. We call these properties **state parameters** or **state functions**, as they describe the state that the system is in at a given time. Note that the value of the state parameters does not depend on how the system gets to a particular state. You are all familiar with some of these state parameters. To list a few: temperature, volume, mass, pressure, concentration, etc.

State parameters are classified (Figure 2.2) into two categories:

- **Extensive parameters** depend on the size of the system or the amount of substance in a system. Examples: mass, volume, number of moles;
- **Intensive parameters** do not depend on the size or the amount of substance in the system. Examples are: temperature, pressure, density, concentration, refractive index.

1.3. Temperature, pressure, heat and work

In intuitive terms, **temperature** is the physical quantity that refers to how hot or cold a particular system is. However, this simple definition neither explains what temperature results from, nor does

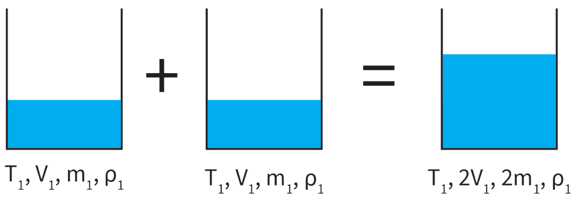


Figure 2.2. Extensive vs intensive parameters. Two identical thermodynamic systems (water in a glass) are mixed, resulting in a thermodynamic system that has the same temperature and density (intensive parameters) as the original systems, but double the mass and volume (extensive parameters).

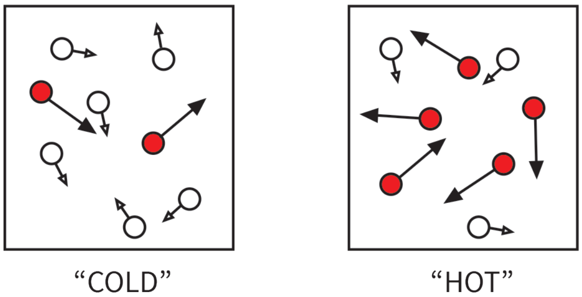


Figure 2.3. Comparison of two simplified thermodynamic systems – boxes holding an equal number of gas molecules that are in motion. Two types of gas molecules are shown – fast (high kinetic energy, full circles) and slow (low kinetic energy, empty circles). The average kinetic energy is higher in the rightmost box, hence the gas in it has a higher temperature than the gas on the left.

it give a quantitative meaning to this parameter. In order to understand temperature better, we have to first discuss the concept of **kinetic energy**.

All objects in motion possess a certain kinetic energy, which is proportional to the squared speed that the objects are moving at. Thus, an object that moves faster has a higher kinetic energy than one that moves slower. The atoms and molecules that make up the world around us are always in motion (called *thermal motion*), and thus possess a non-zero kinetic energy. Even in the case where an object does not appear to be moving at a macroscopic scale (to our eyes), its individual atoms or molecules are still in motion. As not all atoms or molecules in a given object have the same kinetic energy, when talking about a property of the entire object, it makes more sense to talk about the average kinetic energy of all its particles. In a simple definition, we can consider temperature as a measure of the average kinetic energy of the particles in a thermodynamic system (Figure 2.3).

Several temperature scales are in use worldwide. The most commonly used is the **Celsius scale**, which measures temperature in degrees Celsius (°C). The Celsius scale is defined based on two points – the melting point of ice (0 °C) and the boiling point of water (100 °C), both measured at a pressure of 10^5 Pa.

The International System of Units (abbreviated SI, from the French name: Système International) uses the **Kelvin scale**, having the Kelvin (K) as a unit of measurement. The Kelvin scale is defined based on the extrapolated temperature at which all molecular motion stops, which is 0 K. This corresponds to –273.15 °C. **No temperature lower than 0 K can exist.** As the temperature increment in the Kelvin scale is the same as in the Celsius scale (a difference of 1 °C is the same as a difference of 1 K), we can easily convert temperatures between the two scales according to:

T (°C)	T (K)
–273	0
–50	223
0	273
25	298
37	310

Table 2.1. Temperature conversion between the Celsius and Kelvin scales. For convenience, the values are rounded to the nearest integer.

$$T\,(K) = T\,(°C) + 273.15 \quad (2.1)$$

Quite often, the decimals ".15" are neglected so that the temperature is expressed as an integer (Table 2.1).

For fluid (gas or liquid) particles (atoms or molecules) placed in a container, thermal motion such as that presented in Figure 2.3 will result in some particles hitting the walls of the container with a certain force. The ratio between the force exerted on the walls of the container and the total surface of the walls of the container is called the **pressure**. In the International System of Units, pressure is measured in N/m^2, a unit called the pascal (Pa).

If two objects (thermodynamic systems) that have different temperatures are placed in contact with one another, we observe that, over time, the temperature of the hot object will decrease while that of the cold object will increase. After a while, thermal equilibrium has been reached when their two temperatures equalize. We can, therefore, conclude that energy has been transferred from the hot object to the cold object. In thermodynamics, we call this transferred energy **heat**. Heat transfer results from the disordered movement of atoms and molecules. Note that temperature is not the same as heat, although they are often used interchangeably in common language. As **heat is a form of energy transfer**, it is measured with units of energy. In the International System of Units, that unit is the joule (J).

The **zeroth law of thermodynamics** is the one that introduces temperature as a universal physical quantity. In brief, this law says that if a system A is in thermal equilibrium with another system B and B is in thermal equilibrium with a system C, then A and C are also in thermal equilibrium with each other – or, otherwise said, they have the same temperature.

Another commonly used unit of measurement for energy in thermodynamics is the calorie (cal). One calorie is the energy needed to increase the temperature of one gram of water by one degree Celsius (from 14.5 °C to 15.5 °C).

$$1\,cal \approx 4.18\,J \quad (2.2)$$

The multiple of the calorie, the kcal (1 kcal = 1000 cal) is extensively used when expressing energy obtained from food items. Although in common language the kilocalorie is often mistakenly called "calorie" as well, note that it is actually 1000 times larger than the calorie.

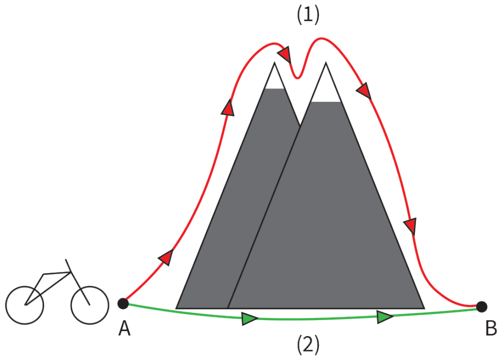


Figure 2.4. Work is not a state parameter. The biker can take either path 1 (up and down the mountains) or path 2 (around the mountains) to get from point A to point B. Taking path 2 uses less energy, or otherwise said less work is performed.

If the transfer of energy performed as a result of disordered movement is called *heat*, what is the term used when energy is transferred as a result of *ordered movement*? This is called *mechanical work* or simply **work**. The physical definition of work (W) is the energy expended when a force (F) is exerted on an object, causing it to move over a certain distance (d). Some other examples of work are: changing the volume of an object against a pressure, changing the surface of an object against an opposing force, electrical charge transport in an electrical field, etc. Thus, by definition:

$$W = F \cdot d \quad (2.3)$$

Note that, strictly speaking, work is a scalar product of two vectors, force and displacement. Equation (2.3) is valid only when these two vectors are parallel, thus the force is exerted parallel to the direction of travel.

As **work is also a form of energy transfer**, it is measured in the same units as heat (joules, calories, etc.). Note that both heat and work depend on the path that a system takes to get to a certain point. Thus, **heat and work are not state parameters** (Figure 2.4).

1.4. Thermodynamic processes

A system that changes its state (its state parameters change) undergoes a **thermodynamic process**. The system moves from an initial state towards a final state by going through a number of intermediate states. If the system is able to return to the initial state by going through the same intermediate states, in reverse order, we say that the process was **reversible**. If the return towards the initial state can only happen through different

intermediates, we call the process **irreversible**. In reality all thermodynamic processes belong to this second category, but it is sometimes convenient to analyze idealized systems, in which some processes can be considered reversible.

1.5. Particular thermodynamic states

A thermodynamic system that reaches a state in which all exchange of energy and substance stops is said to be in a state of **thermodynamic equilibrium**. At equilibrium, all state parameters are constant in time (they stay the same over time) and intensive parameters are also constant in space (they have the same value in all points of the system).

Another particular thermodynamic state is the **steady state**. In this state, intensive state parameters can have different values in different points of the system, but they are constant over time. A steady state can only exist if the system is able to exchange substance or energy with its environment, thus is cannot exist in isolated systems.

The natural tendency of thermodynamic systems is to evolve towards a state of equilibrium. For instance, consider a bathtub filled with warm water. If at one end of the bathtub someone pours cold water, the water temperature is different now in different places of the bathtub. Without stirring the water, we observe that after a while the water temperature equalizes throughout the bathtub; this is an example of the equilibrium state.

Living biological systems are not in a state of equilibrium, as that would mean death. Consider a living warm-blooded organism. The temperature of its body does not change (let say it is 37 °C). This appears to be an equilibrium situation, but it is not. Most state parameters are not constant in space (this goes without saying – a system at equilibrium has no internal organization and looks rather like a “soup” in which all its component atoms and molecules are randomly distributed). The organism breathes, loses heat through perspiration (cools down) but heat is generated through metabolic processes (“burning” of glucose). This situation is the steady state. If the organism dies, the steady state is lost, and the organism will enter equilibrium with the environment: as a simple example, it cools down to the same temperature as the environment.

Steady states are the most interesting in biological thermodynamics. Living organisms have extremely well-organized systems to maintain it. Unfortunately, these steady states are also complicated to study because many state parameters change simultaneously; in order to describe them, knowledge about entropy and fluxes is required (these are discussed later in this chapter).

1.6. The law of ideal gases

In real life, thermodynamic systems can be very complicated as they can contain numerous compartments, different substances, and a multitude of processes happen at the same time. A living cell is an example of an overly-complicated thermodynamic system. To this date, we cannot yet fully describe all the processes and states of a living cell; to understand the basic processes at least, we can study in turn very basic systems, such as molecules of gas enclosed in a compartment. The study of these simple systems gives us an overview of a complicated system.

A gas in which the individual atoms and molecules are considered material points that do not interact with each other in any meaningful way is called an **ideal gas**. Under many conditions, most real gases behave as ideal gases. Between the state parameters of a system containing an ideal gas, the following relation, called the **law of ideal gases** (fundamental law of gases), was found to be true:

$$pV = \nu RT \qquad (2.4)$$

where p is pressure, V is the volume, ν is the number of moles, T is the temperature and R is a constant called the constant of ideal gases ($R \approx 8.31\ J\cdot mol^{-1}\cdot K^{-1}$).

The law of ideal gases shows what will happen in a system if one or more parameters change. Some simple examples:

- If in a system V = constant and T increases → the thermal motion increases → the kinetic energy of particles increases → the average impact force on the walls of the container increases over the same surface → we will observe an increase of the pressure p;
- If in a system T = constant, and V increases → the surface of the walls increases, but the kinetic energy of the molecules is the same → the same impact average force is spread over a bigger area → we will observe a decrease of pressure p.

The law of ideal gases is important in several fields besides basic thermodynamics. For instance, we’ll discuss in a later chapter the very similar law of van't Hoff for osmotic pressure.

Now that we have presented these fundamental notions, we can start the discussion of the **laws of thermodynamics**. These are not “laws” in the legal sense. You can break a law that tells you, for instance, not to cross the street if you see a red traffic light (and suffer the consequences), but you cannot break the laws of thermodynamics – they are principles that were found to be generally true for the physical universe.

2. FIRST LAW OF THERMODYNAMICS

2.1. Internal energy

The 1st law of thermodynamics is a consequence of a more general physical law called the **principle of conservation of energy**. In brief, this principle states that **energy cannot be created from nothing, nor can it dissappear into nothing – it can only be transformed**.

In order to express the 1st law mathematically, a new extensive state parameter called **internal energy** is introduced.

The **internal energy (U)** is the total energy of a system, which includes both the kinetic energy of the system's components (translations, vibrations, rotations) and the potential energy of the system's components (attraction and repulsion between atoms and molecules, chemical bonds, interactions with force fields inside the system, etc.). Note that the internal energy does not include the kinetic energy of the system as a whole (how the system moves relative to its environment), nor does it include the system's potential energy under external force fields.

The first law states that, whenever the internal energy of a system changes, this change must be a result of energy gained or lost by the system. We have seen that, in thermodynamics, energy can be transferred as heat or mechanical work. Thus, a system can gain energy, for example, by absorbing heat and can lose energy by performing mechanical work. Thus, the mathematical formulation of the first law of thermodynamics is:

$$\Delta U = Q - W \qquad (2.5)$$

In equation (2.5), we write as "ΔU" the difference in internal energy between the final state and the initial state. You will also encounter this notation later in this chapter. The transferred heat (Q) is taken as positive if the system absorbs heat and negative if the system loses heat. The work (W) performed by the system is taken to be positive, while work performed on the system is taken as negative.

Remember that U is a state parameter while Q and W are not. The value of U at a given point in time depends only on the measurable properties of the system at that time and ignores how the system arrived at that state. Thus, the variation of internal energy of a system, ΔU, depends only on the final state and initial state of the system.

If we consider an isolated system, which cannot exchange energy with its environment, we immediately see that $Q = 0$ and also $W = 0$. Consequently, $\Delta U = 0$. Thus, **the internal energy of an isolated system stays constant over time**.

Nutritional information	for 100 g
Energy value	536 kcal
Fat	36 g
Carbohydrate	47 g
Protein	6 g

Table 2.2. Nutritional information for a bar of chocolate.

2.2. Application: circuit of energy in the biosphere

The main source of energy for the living world is solar radiation. This is absorbed by plants, which perform *photosynthesis*, converting CO_2 and H_2O to complex molecules (carbohydrates) and releasing O_2. We call plants and other organisms that can produce complex molecules from simple sources of energy such as sunlight *autotrophs*.

Organisms classified as *heterotrophs* (a category which includes, among others, humans) obtain their energy by consuming autotrophs or other heterotrophs.

2.3. Application: calculating energy intake

As mentioned previously, we express the amount of energy gained by consuming a particular food in kcal. Current regulatory standards in the European Union mandate the energetic content of food and drink to be listed on the packaging. Essentially, on the packaging of almost all food or drink that you buy, you will find a table that closely resembles Table 2.2, stating the amount of energy (in kcal or kJ) that you gain from consuming 100 g of solid food or 100 mL of a drink.

We see that eating 100 g of the chocolate in Table 2.2 is equivalent to an energy intake of 536 kcal. How does one reach this number? Studies have shown the average amount of energy that is gained by a person when consuming one gram of a certain nutrient. This is: 4 kcal/g for carbohydrates, 4 kcal/g for proteins and 9 kcal/g for fats.

Therefore, to determine the energy content of a given food, the food is first subjected to chemical analysis. In this way, the nutrient content of the food is determined. After this, the energy value is calculated as:

$$\textit{total energy (kcal)} = g_{ch} \cdot 4 + g_{pr} \cdot 4 + g_f \cdot 9 \qquad (2.6)$$

where g_{ch} = grams of carbohydrate, g_{pr} = grams of protein and g_f = grams of fat.

Note that, without going into details of nutrition and diets, the first law of thermodynamics states a simple condition for losing weight – if energy input is reduced under conditions where a person performs the same amount of activity (work), the internal energy has to decrease (for example, by burning stored fat).

2.4. Application: energy balance in the body and calculating the basal metabolic rate

The human body is a thermodynamic system and thus follows the 1st law of thermodynamics, or otherwise said the energy intake of a living organism has to be equal to the energy output. We can write this energy balance in the form of the following equation:

$$\textit{Energy intake} = \textit{Energy expenditure} + \textit{Stored energy} \tag{2.7}$$

In a healthy adult that maintains a stable body weight, the energy intake gained by consuming food is equal to the energy they expend. If the energy intake increases but the expenditure stays the same, the excess is stored as fat. The energy expenditure of the body can be detailed as:

$$\textit{Energy expenditure} = \textit{Physical activity} + TEF + BMR \tag{2.8}$$

where TEF = thermic effect of food and BMR = basal metabolic rate.

Let us discuss the terms of equation (2.8) in turn. The energy expended by physical activity is, for a normal individual, about 25% of their total energy expenditure. The basal metabolic rate (BMR) represents the minimum energy expenditure needed for an organism to survive (the energy they expend at rest). BMR can account to up to 70% of a person's energy expenditure. The thermic effect of food (TEF) represents the increase of the metabolic rate above the BMR associated with digestion, absorption and storage of food in the body. The TEF is different depending on the types of food consumed: it is much higher for proteins than for fat or carbohydrates.

Measuring BMR allows the determination of a patient's nutritional needs and offers a baseline for establishing a diet. The most accurate method of determining BMR is by direct calorimetry. This involves confining the patient in a special room (whole-room calorimeter) and measuring the heat exchanged by the body with the environment. As this technique requires highly specialized equipment that is only available in few facilities around the world, it is rarely used.

Instead, BMR is often determined by indirect calorimetry (respirometry). This technique uses a ventilated hood that is placed over a patient's head and is made airtight around the neck. The amounts of O_2 used and CO_2 produced by the patient at rest are used as an estimate of the heat produced (hence the term indirect). Measuring BMR this way requires that some strict conditions are met beforehand: the patient must not have eaten for at least 12 hours before the measurement, must have slept for 8 hours and must be at rest during the measurement. If these strict requirements are not met, the amount of heat measured is termed the resting metabolic rate (RMR), which is slightly higher than the BMR.

Due to the cost and time required for measuring BMR, its value is, instead, usually estimated by the use of various empirical equations.

An example of one such equation, the Mifflin – St. Jeor equation, is:

$$BMR = 10W + 6.25H - 5A + S \tag{2.9}$$

where W is the weight of the individual in kg, H is the height in cm, A is the age in years and S = +5 for men and −161 for women.

3. SECOND LAW OF THERMODYNAMICS

3.1. Entropy

The first law of thermodynamics sets a clear limitation on thermodynamic processes: you cannot have energy disappear into nothing, nor can you create it from nothing. However, the first law tells us nothing about thermodynamic processes having a preferred direction. Nevertheless, this is an observation that we have all made in real life.

Consider the following example (Figure 2.5) of a glass of water in which a drop of dye is added. This is a simple experiment that you could also do at home. What happens to the thermodynamic system in the absence of any outside intervention? We observe over time that the dye distributes itself in the water. Indeed, if we leave the glass sitting long enough, we will not be able to tell where we added the drop initially – the entire water will be stained by the dye. No matter how many times we repeat this experiment, we will observe the same thing – the drop of dye disappears. We have seen that this thermodynamic system spontaneously and irreversibly evolved from an ordered state (all the molecules of dye concentrated in one drop) towards a state of maximum disorder (the molecules of dye are randomly distributed in the water).

In thermodynamics, we express disorder by an

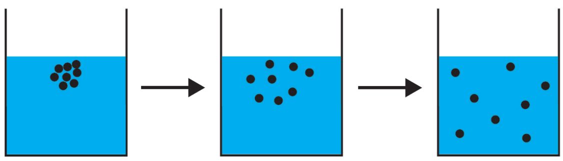


Figure 2.5. A drop of dye is added into a glass of water. Without any outside intervention, we observe over time that the dye will distribute itself in the entire volume of water.

extensive state parameter called **entropy (S)**. The higher the degree of disorder in a system is, the higher the value of the entropy is.

We can now state one form of the second law of thermodynamics: "**in isolated systems, every spontaneous process is one in which entropy increases**". In thermodynamics, the word **spontaneous** defines a process that occurs without needing external input of energy. The opposite of a spontaneous process is a **non-spontaneous** one, which only occurs if external energy input is provided. Obviously, as isolated systems cannot exchange energy with their environment, non-spontaneous processes cannot happen in those systems. Note that the second law does not say anything about the speed at which a process takes place, which is a matter studied by another branch of science called kinetics.

Another formulation of the 2nd law is that of Clausius, who stated: "**heat can never pass from a colder to a warmer body without some other change, connected therewith, occurring at the same time**". Indeed, by placing a hot object in contact with a cold one, we will always observe the same spontaneous process – the cold object will get warmer and the hot object will get colder. That does not mean that it is impossible to cool down a hot object while transferring heat to a warmer area. Indeed, this is what our refrigerators and air conditioning units do all the time. However, in order to do that, these machines expend work and expend energy from an external source (the power grid).

We can mathematically define the change in entropy (ΔS) of a system undergoing a reversible process as:

$$\Delta S = \frac{\Delta Q}{T} \tag{2.10}$$

where ΔQ is the heat transferred at a certain temperature T.

Note that the way entropy is defined makes it obvious that you generate more entropy by heating up a colder system than a warmer one.

For $\Delta S > 0$, the entropy of the system increases, while for $\Delta S < 0$, the entropy of the system decreases (remember this cannot happen in isolated systems). If $\Delta S = 0$, the process is reversible and can occur in either direction (this is an idealized case, as stated previously, in real life processes are irreversible).

There is also a **third law of thermodynamics**, which defines an absolute value of entropy. The third law states that the entropy of an object is minimum when temperature is the absolute zero (0 K). For a perfect crystal, this minimal value of the entropy at 0 K is zero.

Table 2.3. Comparison of the equilibrium and steady state.

	Thermodynamic equilibrium	Steady state
In isolated systems?	Yes	No
State parameters	Constant in time	Constant in time
Intensive parameters	Constant in space	Not constant in space (gradients)
Entropy production	Zero	Minimum, but not zero

3.2. Application: thermodynamic equilibrium and the steady state

We introduced in a previous section the concepts of *thermodynamic equilibrium* and the *steady state*. In the state of equilibrium, state parameters are constant over space and constant in time. Therefore, in an isolated system in the state of equilibrium, entropy has reached a maximum value and cannot increase any further. Otherwise said, entropy production is zero at equilibrium.

A system at steady state is not allowed to evolve into a state of equilibrium by the intervention of external forces. We can say that, in such a system, the production of entropy is minimum, but not zero. In order for the system to remain in the steady state, this produced entropy must be eliminated by exchanges with the environment.

Table 2.3 shows a comparison between the thermal equilibrium and the steady state.

3.3. Application: living organisms and entropy

Biological processes are irreversible thermodynamic processes. Thus, they result in an increase of entropy. However, living organisms are highly organized thermodynamic systems. How can we explain this apparent contradiction?

We know that entropy has to increase in spontaneous processes that happen in isolated systems. However, a living being is an open system. Indeed, any living being increases its degree of organization (decreases its entropy) by increasing the entropy of its surrounding environment. Living organisms eat food containing low entropy, complex molecules (proteins, carbohydrates, etc.), that they use either as an energy source or for building and repairing tissues, while excreting simple molecules (such as CO_2 or H_2O) with a high entropy. Thus, even if locally (in the body), the entropy decreases, overall (in the environment), entropy increases.

We can thus conclude that processes which are normally non-spontaneous occur as long as they are coupled to other processes which are spontaneous. Overall, the sum of these processes

(entropy is additive!) should lead to an increase in entropy. We will describe such a case in more detail towards the end of this chapter.

4. THERMODYNAMIC POTENTIALS

In physics, a *potential* denotes the capacity of a system to perform a certain action. For example, in mechanics, the *potential energy* denotes the capacity of a system for performing mechanical work. In the same manner, one can define **thermodynamic potentials**, or, more accurately, **thermodynamic potential energies**, as all the quantities defined here have the physical meaning of energies.

Thermodynamic potentials are extensive state functions, and we have already discussed one of them so far – the internal energy (U). Referring to the definition of U in equation (2.5) we can say that the internal energy is the capacity of the system of performing work plus the capacity of the system to release heat.

4.1. Enthalpy

Most biological systems function under conditions of constant pressure and temperature. Under those conditions, we can calculate the work as $p \cdot \Delta V$ and, according to equation (2.5), the heat transferred can be defined as:

$$Q = \Delta U + p\Delta V = \Delta H \qquad (2.11)$$

where H is the *enthalpy*.

While the absolute value of the enthalpy in a system is hard to determine, the variation of enthalpy is commonly used to characterize chemical reactions (this is also called heat of reaction):

$$\Delta H = \Delta U + p\Delta V \qquad (2.12)$$

Thus:

- If $\Delta H < 0$, the reaction results in the release of heat into the environment. We call this an **exothermic reaction**. A reaction between an acid and a base (neutralization) is an example of such a reaction.
- If $\Delta H > 0$, the reaction occurs with absorption of heat from the environment. We call this an **endothermic reaction**. Dissolution of table salt (NaCl) in water is an example of such a reaction.

4.2. Helmholtz free energy

The second law of thermodynamics states that not all the internal energy of a system can be used to perform work, as some of it is irreversibly degraded into heat to increase entropy. Under conditions of constant temperature and volume, we can define the part of the internal energy that can be used to perform work as the *Helmholtz free energy* (F):

$$\Delta F = \Delta U - Q_{degraded} = \Delta U - T\Delta S \qquad (2.13)$$

where ΔS is the change in entropy, thus $T\Delta S$ is the energy lost as heat to increase the entropy of the system.

4.3. Gibbs free energy

In biological systems, as well as in most common chemical reactions, which occur at constant pressure, the Helmholtz free energy is not a useful quantity to calculate, as it refers to systems at constant volume. Instead, the most important thermodynamic potential in these systems is the *Gibbs free energy* (G), which expresses the maximum amount of work that can be performed by a system at constant temperature and pressure:

$$\Delta G = \Delta U - T\Delta S + p\Delta V \qquad (2.14)$$

The variation of the Gibbs free energy (ΔG) can be used to predict the direction towards which a system will spontaneously evolve at constant pressure and temperature. Using equation (2.12), we can easily see that:

$$\Delta G = \Delta H - T\Delta S \qquad (2.15)$$

For chemical reactions, the value of ΔG shows whether a reaction is thermodynamically favored (is spontaneous) or not. Take for instance, the reaction below, where we wrote a double arrow to indicate that the reaction could theoretically occur in either direction:

$$A + B \rightleftarrows C$$

We can calculate the variation of the Gibbs free energy in this reaction as:

$$\Delta G = G_{products} - G_{reactants} = G_C - (G_A + G_B) \qquad (2.16)$$

There are three possibilities for this reaction:

- if $\Delta G < 0$, the chemical reaction is spontaneous and will occur towards the formation of C (from left to right) without needing external energy input. We call such a reaction **exergonic**. Note that if a reaction is both exothermic ($\Delta H < 0$) and entropically favored ($\Delta S > 0$), it will certainly be exergonic;
- if $\Delta G > 0$, the chemical reaction is non-spontaneous. Instead, the reaction favors the formation

of reactants A and B (the reaction normally occurs from right to left). The reaction will only occur towards the formation of C (from left to right) if external energy is provided. We call such a reaction **endergonic**. Note that a reaction is certainly endergonic if $\Delta H > 0$ and $\Delta S < 0$;

- if $\Delta G = 0$, the system is in equilibrium and the concentrations of the reactants (A and B) and the product (C) will remain constant.

4.4. Application: energy balance of glucose oxidation

Living organisms require input of energy for generating mechanical work in muscle contraction and cell movement, synthesizing molecules and active transport of chemical species. Most of the energy that our body uses results from the oxidation ("burning") of the energy-rich glucose molecule, a reaction which can be written as:

$$C_6H_{12}O_6 + 6O_2 \rightarrow 6CO_2 + 6H_2O \qquad \Delta G = -2870 \text{ kJ/mol}$$

In many cases, release of such a high amount of energy would be, at best, wasteful (much of this energy could not be used for work and would be lost as heat), or, at worst, harmful (it would cause the temperature in the cell to increase locally to dangerous values). Furthermore, consider that, even though this reaction is thermodynamically favored (spontaneous), it does not occur easily at normal temperatures (sugars don't burn by themselves when stored in your kitchen). This is because this direct reaction requires a high activation energy (initial heat provided) in order to occur.

To eliminate all these disadvantages, the biochemical oxidation of glucose is performed in the body stepwise, through a large number of chemical reactions (metabolic pathways). In turn, the energy released in each of these reactions is stored in the form of ATP, the energy "currency" of the body. Overall, no net energy is lost compared to the direct oxidation reaction as shown in Figure 2.6.

4.5. Application: energetic coupling of reactions

Many of the chemical reactions in our body are energetically unfavorable ($\Delta G > 0$). In order for these endergonic reactions to occur, energy needs to be provided by a second, exergonic reaction, coupled to the first. This second reaction is many times the hydrolysis of ATP, which is coupled to the first reaction by an enzyme. An example of such coupling is represented below.

The first step of glycolysis is the conversion of glucose to glucose-6-phosphate, a reaction which is endergonic, thus non-spontaneous:

$$glucose + P_i \rightarrow glucose\text{-}6\text{-}phosphate + H_2O \qquad \Delta G_{(1)} = +14 \text{ kJ/mol}$$

where P_i is the inorganic phosphate ion (PO_4^{3-}).

In order for this reaction to occur, it is *coupled* to the reaction of ATP hydrolysis by the enzyme hexokinase. ATP hydrolysis can be written as:

$$ATP + H_2O \rightarrow ADP + P_i \qquad \Delta G_{(2)} = -30 \text{ kJ/mol}$$

As the two reactions are coupled, we can write the sum of the two reactions and add their ΔG values:

$$glucose + ATP \rightarrow glucose\text{-}6\text{-}phosphate + ADP \qquad \Delta G_{(3)} = \Delta G_{(1)} + \Delta G_{(2)} = -16 \text{ kJ/mol}$$

The new reaction is, thus, spontaneous and, if ATP is available, glucose will be phosphorylated in the presence of hexokinase.

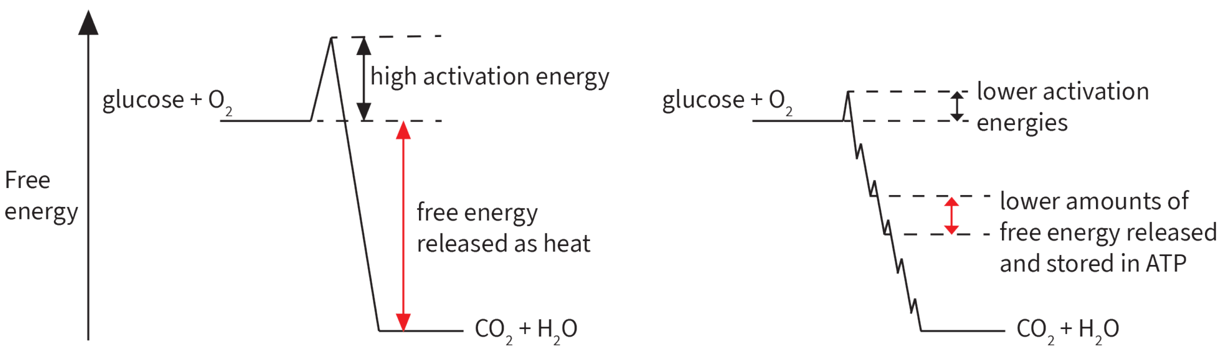


Figure 2.6. Oxidation of glucose can occur directly, in one step (left panel) or in several smaller steps (right panel). The same free energy is released overall in either case, as free energy is a state function and the reactants and end products are the same. However, when stepwise reactions are used, two main advantages are apparent: 1) each individual reaction requires lower activation energies (a process aided by enzymes) and 2) the smaller amounts of energy released in each reaction can be stored in the form of ATP molecules.

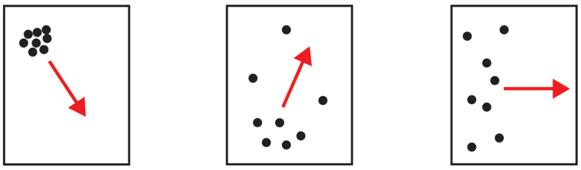


Figure 2.7. Different concentration gradients resulting from the distribution of gas molecules in a container. The arrows indicate the direction of the resulting fluxes.

5. THERMODYNAMIC GRADIENTS AND FLUXES

We have mentioned that, in a system that is not at thermodynamic equilibrium, intensive state parameters can vary in space. We call such a variation a **thermodynamic gradient**.

Mathematically, we define a gradient (X) as the variation of an intensive state parameter over a distance x. As examples:

- $X_T = \dfrac{\Delta T}{\Delta x}$ is a temperature gradient;
- $X_c = \dfrac{\Delta c}{\Delta x}$ is a concentration gradient.

In thermodynamics, gradients play the role of thermodynamic forces – the existence of a gradient causes the appearance of a thermodynamic flux, which represents the change over time and a given surface of a certain state parameter. Examples are given in Figure 2.7.

The observation that the existence of a gradient generates a corresponding flux is nothing more than a consequence of the 2^{nd} law of thermodynamics: systems tend to evolve towards a state of maximum entropy (high disorder = high uniformity = no variation in state parameters).

Writing J as the flux and X as the gradient, it was found that the resulting flux J is proportional to the gradient that generates it:

$$J = LX \tag{2.17}$$

where L is a proportionality constant.

Note that, in physics, the direction of the gradient is mathematically defined as from the lowest towards the highest value. For instance, for the concentration gradients depicted in Figure 2.7, this would be the opposite direction of the arrows (the opposite of the direction of the resulting flux).

In biology, biophysics and biochemistry, quite often the direction of the gradient is taken as the reverse of its physical direction (from high to low). We will generally use this notation. Thus, we will say, for instance, that a solute is transported by diffusion "in the direction of" (or "down") the concentration gradient, meaning from high concentration to low concentration.

REFERENCES

Atkins, P. W., De Paula, J., & Keeler, J. (2017). *Atkins' Physical Chemistry*. London: Oxford University Press.

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

California Institute of Technology, Gottlieb, M. A., & Pfeiffer, R. (2013). *The Feynman Lectures on Physics*. Retrieved from https://www.feynmanlectures.caltech.edu/

Coulston, A. M., Rock, C. L., & Monsen, E. R. (2001). *Nutrition in the prevention and treatment of disease*. San Diego: Academic Press.

Cromwell, B. (2010). *Light and Matter*. Retrieved from http://www.lightandmatter.com/lm/

Flowers, P., Theopold, K., Langley, R., & Robinson, W. R. (2019). *Chemistry 2e*. Retrieved from https://openstax.org/books/chemistry-2e/pages/1-introduction

Franklin, K., Muir, P., Scott, T., & Yates, P. (2019). *Introduction to Biological Physics for the Health and Life Sciences*: Wiley.

Kondepudi, D., & Prigogine, I. (1998). *Modern Thermodynamics. From Heat Engines to Dissipative Structures*. Chichester: John Wiley & Sons.

CHAPTER 3

WATER IN BIOLOGICAL SYSTEMS

Prerequisite knowledge

- Matter phase changes
- Chemical bonds (ionic, covalent, coordinative) between atoms
- Electronegativity of a chemical element

Water is the most important substance for the existence of life on Earth. Due to its structure, water has unique properties, vital for the complexity of living organisms. At least 50 % of the content of living organisms is water and the human body contains 65 – 70 % water. It is the medium for dissolving nutrients (in the extracellular fluids) but also for complex biochemical reactions (inside the cells). At normal body temperature, water is liquid, a state more dynamic compared to the solid state and more complex than the gaseous state.

1. THE LIQUID STATE. INTERMOLECULAR FORCES

1.1. The liquid state

Pure samples of chemically simple substances can usually be found in one of three different physical states: solid, liquid or gas (other states can also exist under more extreme conditions).

A solid has a definite volume, density and shape because the intermolecular forces between its molecules are so strong that the molecules are not allowed to change their positions. However, molecules in a solid are not completely still as they vibrate about their fixed positions.

In a liquid, molecules are bound strongly enough to give the liquid a fixed volume and density, but there is also a higher degree of molecular movements, because of which liquids do not have a definite shape. The thermal kinetic energy of a liquid molecule is high enough to break the intermolecular bonds, but not to completely escape from its neighboring molecules. Thus, the molecules in a liquid can move relative to the others but within a fixed volume.

In the gas state, the molecules are free to move away from each other, therefore the volume and shape are defined only by the container they're in.

1.2. Polar and non-polar molecules. Dipoles

Depending on how electrical charges are distributed in a molecule, we say that some molecules are *polar* (also called *dipoles*) while other molecules are *non-polar*.

Let's see what that means. Consider a molecule formed by two identical atoms bound by a covalent bond such as F_2 (Figure 3.1). In this situation, the two atoms have the same electronegativity[1] and share their bonding electrons evenly, so there is no separation of charge. We call such a molecule **non-polar**.

If we replace one of the fluorine atoms with hydrogen we obtain hydrogen fluoride (HF, Figure 3.1). As in the case of F_2, this molecule is still electrically neutral. However, the two atoms now have different electronegativities: F is much more electronegative than H. Because of that, the bonding electrons will be shared unevenly: the electrons spend more time around F than around H. For this reason, the fluorine side of the bond is slightly negative and the hydrogen side of the bond is slightly positive. We call the HF molecule a polar molecule as it has two electrically distinct sides, or poles: a positive pole (the hydrogen atom) and a negative pole (the fluorine atom). Polar molecules are also called **permanent dipoles**, or simply **dipoles**.

We say that the two sides of the dipole have partial charges, which can be represented by the Greek letter δ (e.g. $^{\delta+}HF^{\delta-}$, Figure 3.1) or by a

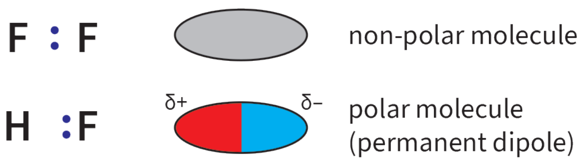


Figure 3.1. Comparison between the HF and F_2 molecules. Both molecules are kept together by a single covalent bond, formed by each atom in the molecule sharing one of its electrons. In F_2, no separation of charges exists and the molecule is non-polar. In the HF molecule, electrons will be more attracted to the fluorine atom, leading to an uneven charge distribution – the molecule is polar (a permanent dipole). Blue dots represent the electrons forming the covalent bond.

1 *Electronegativity* is the tendency of an atom to attract shared electrons in a chemical bond.

crossed arrow oriented toward the negative side of the dipole.

If we think now about molecules that have more than one chemical bond, (therefore containing more than two atoms), assessing the polarity becomes more complex, as this depends also on the overall geometry of the molecule. For example, in the carbon dioxide molecule (CO_2), oxygen is more electronegative than carbon, but the attraction of electrons by one O atom is balanced by the attraction of the other O, as the molecule is linear (the angle between the two C=O bonds is 180°). Therefore, there is no net dipole and CO_2 is a non-polar molecule. However, there are many examples in which the dipoles of different bonds do not cancel each other and the molecule as a whole is a permanent dipole: water is an example of such a molecule. There are also molecules which contain polar groups as well as non-polar groups and these are called amphiphilic (or amphipathic), for example lipids.

Non-polar molecules can temporarily become dipoles, forming *instantaneous dipoles* or *induced dipoles* (Figure 3.2).

An **instantaneous dipole** is a short-lived species that forms spontaneously as a consequence of the constant movement of electrons, which can cause the charge distribution in a molecule to briefly become asymmetrical.

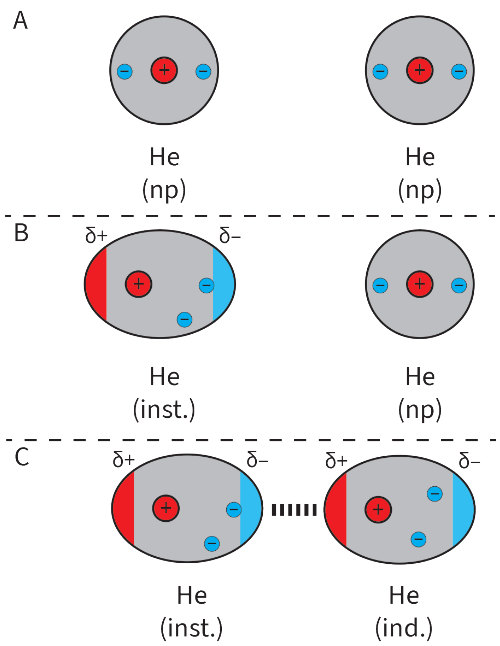


Figure 3.2. Instantaneous and induced dipoles. A model system with two helium (He) atoms is used. The He atom has two protons (in the nucleus) and two electrons. A, symmetrical distribution of charges in both He atoms makes them non-polar. B, the leftmost He atom randomly becomes an instantaneous dipole through the movement of its electrons. C, the newly formed instantaneous dipole polarizes the rightmost He atom, turning it into an induced dipole. An attractive force (van der Waals interaction) will appear between the two atoms (see below). np = non-polar, inst. = instantaneous, ind. = induced.

When an instantaneous dipole or a permanent dipole come in close proximity to a non-polar molecule, they can alter the charge distribution in the non-polar molecule, transforming it into an **induced dipole**. A molecule or atom that becomes an induced dipole is said to be *polarized*.

1.3. Intermolecular forces. Van der Waals forces

You should be familiar with the notion of **chemical bonds**, such as covalent bonds or ionic bonds. These are strong interactions that keep different atoms together. For example, the fluorine molecule F_2 (Figure 3.1) is formed by two fluorine atoms bound by one covalent bond. This bond is formed by each of the fluorine atoms sharing one outer electron. In order to break the covalent bond between these atoms, a large amount of energy (~150 kJ/mol) has to be provided. As a general rule: **the more energy is required to break a bond, the stronger the bond is.**

Unlike covalent bonds or ionic bonds, **intermolecular forces** are much weaker (require much smaller energies to break), and are short-lived. Intermolecular forces arise between different molecules or regions of the same macromolecule and are the result of electrostatic attraction and repulsion. The intermolecular forces present in liquids can be:
- van der Waals forces;
- hydrogen bonds;
- ion-dipole interactions.

The van der Waals forces (or interactions) include a number of different intermolecular interactions and they are among the weakest intermolecular forces (bonding energies of 1 – 10 kJ/mol). Van der Waals interactions are generally attractive (but they can also be repulsive) forces between molecular components, whole neutral molecules, noble gas atoms, or supramolecular particles without the formation of chemical bonds. These forces appear as consequence of the variation of electrical charge distribution in molecules. **Thus, van der Waals interactions are electrical in nature**. Because of this, they are non-directional (do not rely on a particular orientation of the atoms or molecules).

There are different types of van der Waals interactions (Figure 3.3) depending on the types of atoms or molecules involved; these were named in honor of the scientists which described them. In the increasing order of their strength, they are:
- **London (dispersion) forces** appear between two instantaneous dipoles or between an instantaneous dipole and an induced dipole. London forces are universal and can appear between all types of molecules. Their strength increases with

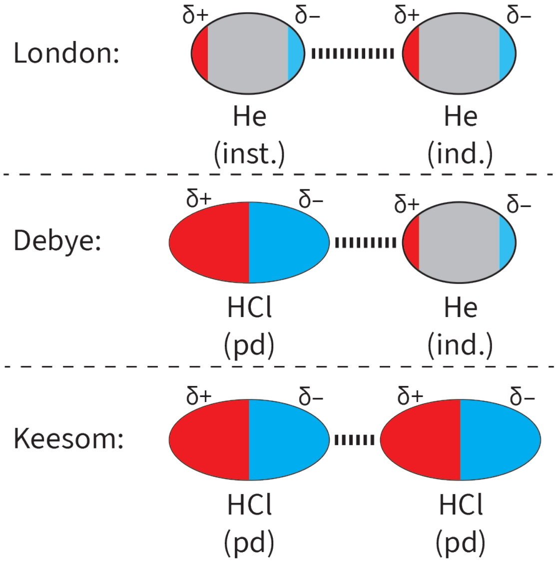


Figure 3.3. Types of van der Waals interactions. Description is provided in the text. pd = permanent dipole; np = non-polar; inst. = instantaneous dipole; ind. = induced dipole.

the molecular mass;

- **Debye (induction) forces**: appear between a permanent dipole and an induced dipole;
- **Keesom (orientation) forces**: appear between two permanent dipoles.

London forces are actually the most important out of all van der Waals interactions. Even though they are individually weaker than the other van der Waals interactions, in most molecules they have the main contribution to the overall intermolecular interactions.

Other intermolecular forces in liquids are ion-dipole interactions. The greater the polarity of a molecule, the stronger the attraction to its neighboring molecules. The ion-dipole forces are stronger (bonding energies of 10 – 20 kJ/mol) than van der Waals forces, which are the weakest intermolecular interactions. However, the strength of even the strongest of these attractions is much weaker than any chemical bond (covalent or ionic bond).

1.4. The hydrogen bond (H-bond)

The *hydrogen bond* is a particular case of an unusually strong dipole-dipole interaction. **The hydrogen bond is the attraction between a hydrogen atom bound to a highly electronegative atom belonging to a certain molecule (usually F, O or N) and a highly electronegative atom belonging to another molecule** (Figure 3.4).

Let us consider two different molecules containing highly electronegative atoms that we will call A and B. Additionally, in the first molecule, a hydrogen atom is covalently bound to A. We can write the H-bond that will be formed between the

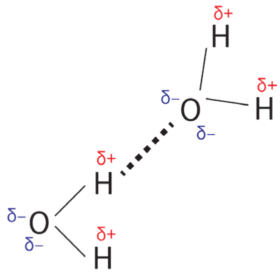


Figure 3.4. A hydrogen bond (dashed line) can be formed between a hydrogen atom of one water molecule and the oxygen atom of a different water molecule. δ+ and δ− denote the partial charges in the water molecule dipole. The overall arrangement in this figure is a water dimer – two water molecules attracted to each other by a single hydrogen bond.

two molecules as: **A–H···B**, where the symbol "···" represents the H-bond.

The hydrogen bond is mainly an electrostatic interaction. The hydrogen side of a polar molecule (like water) is slightly positive because the more electronegative atom A (oxygen, in the case of water) attracts the electrons of the covalent bond more than the hydrogen. As a consequence of that, the hydrogen is electrically attracted to the pair of nonbonding electrons of the negatively charged atom B of the other molecule (another oxygen atom, in the case of water).

Due to the small size of the hydrogen atom and its participation in a covalent bond with another highly electronegative atom, a second electronegative atom can come very close to the hydrogen, without being repelled by its electron shell. As such, the hydrogen bond is mainly electrical, but has some characteristics of a covalent bond, putting its strength in between these two types of interactions. Compare, for instance, the energies required to break the following bonds: ~460 kJ/mol for the O–H covalent bond and ~21 kJ/mol for the H-bond in water, while van der Waals interactions generally have dissociation energies of at most 10 kJ/mol. Thus, **the hydrogen bond is stronger than van der Waals interactions, but much weaker than covalent or ionic bonds**. Additionally, due to its partially covalent nature, the H-bond is directional (requires a particular alignment of the two molcules), unlike van der Waals interactions.

In large molecules such as proteins or nucleic acids, hydrogen bonds can also appear between different regions of the same molecule. We call those intramolecular hydrogen bonds. The rule of formation is the same: a hydrogen atom bound to a highly electronegative atom is attracted to another highly electronegative atom.

Even though the H-bonds are much weaker than any covalent or ionic bonds, the effects of hydrogen bonding can be very pronounced. H-bonds are very important for the dynamics of physiological systems. Because their energy is comparable to random thermal energy, H-bonds can be broken under physiological conditions

(unlike covalent bonds), so they have a short life time (about 10^{-11} s for water molecules). H-bonds also contribute to the structures of organic molecules, via intramolecular interactions, being part of the internal structure of the molecule (e.g. protein, DNA), or contributing to the intermolecular interactions within a group of molecules (e.g. water). They are an important factor in the chemistry of nucleic acids and proteins. The DNA molecule, for example, has a large number of H-bonds and, although the energy of each bond is on the same order of magnitude as that of water H-bonds, the high number of H-bonds (two or three for every base pair), makes the double stranded DNA very stable.

In addition to that, many properties of water derive from H-bonds. One of these is the boiling point. In general, the boiling point of a substance is a measure of the intermolecular forces and it increases with molecular weight due to the increasing strength of the London forces.

Water has an unusually high boiling point (100 °C at normal pressure) compared to similar compounds. For example, at normal pressure, H_2S boils at –61 °C although it has a higher molecular weight than water. This can only be explained by the high amount of energy required to break the H-bonds in order to pass from the liquid to the vapor state. In conclusion, the ability of water molecules to form H-bonds is the reason for which water is liquid at room temperature.

Based on the types of intermolecular interactions between their molecules, we can classify liquids as:

- *simple liquids*, in which van der Waals forces between instantaneous dipoles are the only intermolecular interactions (e.g. halogens, liquefied noble gases like He, Ne, etc.);
- *complex liquids*, in which the molecules attract each other through multiple types of forces, including H-bonds (e.g. water, ethanol, etc.).

2. STRUCTURE AND PROPERTIES OF WATER

2.1. The water molecule (H_2O)

Inside the water molecule, the oxygen atom is covalently bound to the two hydrogen atoms: the distance between the O atom and the H atom is ~0.97 Å (angstrom[2]), and the angle between the two O–H bonds is 104.5°.

In order to have a visual image regarding the spatial arrangement of the atoms in a molecule, and later to understand the structure of liquid water, we also need to take into consideration the

2 1 Å (angstrom) = 0.1 nm = $1 \cdot 10^{-10}$ m.

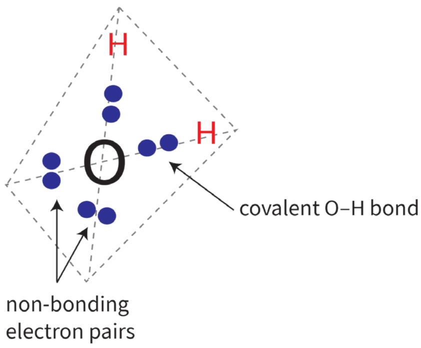


Figure 3.5. Tetrahedral spatial arrangement of the water molecule. Blue dots represent the valence electrons of the oxygen and hydrogen atoms. Each hydrogen atom contributes its single electron to a covalent bond; the oxygen atom also contributes one of its valence electrons to each of these bonds. The remaining 4 valence electrons of oxygen are non-bonding.

two pairs of non-bonding electrons of the oxygen atom. Where do these come from? The oxygen atom has 6 valence electrons (6 electrons in its outer electron layer) and it will use 2 of them to form 2 covalent bonds with the hydrogen atoms, while each hydrogen shares its single electron for these bonds. The remaining 4 valence electrons form two non-bonding electron pairs. These electrons pairs repel each other and the electrons from the hydrogen atoms and, because of that, the geometry of a water molecule is that of a tetrahedron (a triangular pyramid), with oxygen in the center, while the hydrogen atoms and the two pairs of non-bonding electrons are positioned at the corners (Figure 3.5).

If the water molecule were a perfect tetrahedron, the angle between the O–H bonds would be 109.5°. It is actually smaller, at 104.5°, due to the slightly stronger repulsion of the non-bonding electron pairs for each other than for the hydrogen atoms.

The electrons in the covalent O–H bonds are not equally attracted by the oxygen and hydrogen nuclei. They are more attracted by the oxygen atom causing a slightly positive charge around the hydrogen nuclei and a slightly negative charge around the oxygen nucleus, **making the water molecule a permanent dipole** (Figure 3.4).

2.2. The structure of water

Because of its geometry, **a single water molecule can bind 1 – 4 other water molecules** via H-bonds. The most stable arrangement is the water dimer (two water molecules are connected through a hydrogen bond), where one hydrogen atom is attracted to the oxygen of the other molecule (Figure 3.4).

When water is in the liquid state, the H-bonds have a very short life time (around 10^{-12} – 10^{-9}s); they can be broken by the thermal agitation of

molecules and, because of that, the structure of water is not completely understood and still under study. There are various models meant to explain its properties.

The very high number of H-bonds continuously forming in liquid water determines the formation of transient, ordered multimolecular water clusters that diminish with increasing temperature. One possible arrangement of a cluster is a tetrahedral one, in which a water molecule (the core) forms the maximum number of H-bonds connecting with other four water molecules, positioned around the four corners of the tetrahedron (Figure 3.6). The clusters also have short lifetimes. Every H atom will form 100 – 1000 times an H-bond with the same oxygen atom before it connects to another one, and the term flickering clusters is used because of these dynamics. This model was proposed in the 1950's, but later it was concluded that it cannot fully explain the properties of liquid water.

According to a more recent model, on a very short period of time (~10^{-9} s), water behaves like a gel, composed of a single high dimension cluster, which can have local discontinuities and its size depends on temperature and pressure. Due to breaking and reforming of the H-bonds, the cluster changes its configuration, but apparently the formation of more stable polymeric clusters, $(H_2O)_n$, is also possible in certain areas. Various theoretical studies suggest that the size of such a cluster contains on average 90 water molecules at 0 °C and at most 25 molecules at 70 °C. In addition, the existence of an icosahedral cluster composed of 280 water molecules was predicted which could explain the unusual properties of water. However, at the moment, there is no consensus regarding the structure of liquid water; the instability of this kind of multimolecular structures makes their experimental detection very difficult.

When water is in the solid phase (ice), the energy of the H-bonds becomes stronger than the kinetic energy of the water molecules. Thus, the H-bonds become permanent and it is possible for the water molecules to form stable ordered structures (crystalline networks). In these crystalline ice structures, water molecules are connected via the maximum number of H-bonds (each molecule binds four other molecules). These structures can have various shapes, depending on temperature, pressure and the presence of impurities.

Depending on various conditions, several ice structures were found (up to now, 19 such structures, also called *phases*, have been identified). The basis for all of them is the tetrahedral structure, in which each oxygen atom sits at the vertex of a tetrahedral network of bonds (two covalent bonds with the hydrogen atoms from the same molecule and two hydrogen bonds with hydrogen atoms of other molecules). With increasing number of water molecules connected this way, a three-dimensional network is built up leading to the crystalline structure of ice.

Experiments of X-ray diffraction have shown that, at normal pressure and temperatures from –80 °C to 0 °C, water molecules naturally arrange themselves into a structure called **hexagonal ice**. This is, by far, the most common structure of ice in nature. Due to the high number of H-bonds present in the solid state, the positions of water molecules are at the corners of hexagonal prisms. As Figure 3.7 shows, this type of arrangement creates a rather open network, with lots of space between the molecules (more space than in the

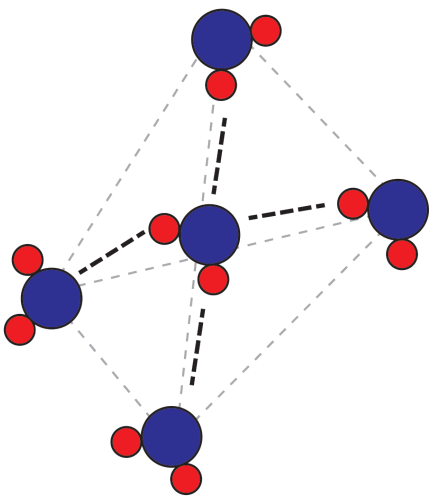


Figure 3.6. Spatial arrangement of a cluster of four water molecules. One water molecule (center) forms four hydrogen bonds (black, dashed lines) with other water molecules. Grey dashed lines show the sides of the tetrahedron, but have no physical meaning.

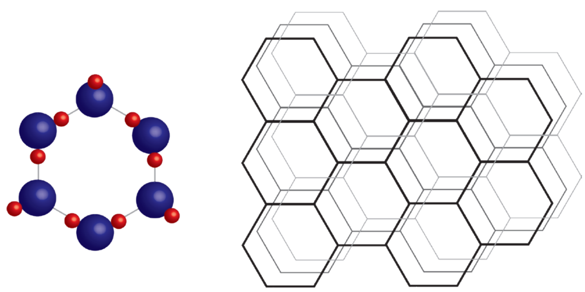


Figure 3.7. Hexagonal ice. Left: 6 water molecules arranged in a hexagonal structure are bound by hydrogen bonds. Large circles = oxygen atoms; small circles = hydrogen atoms. As the structure is 3-dimensional, some hydrogen atoms are behind the oxygen and will form hydrogen bonds with other hexagonal planes. Right: 3-dimensional network of hexagonal ice.

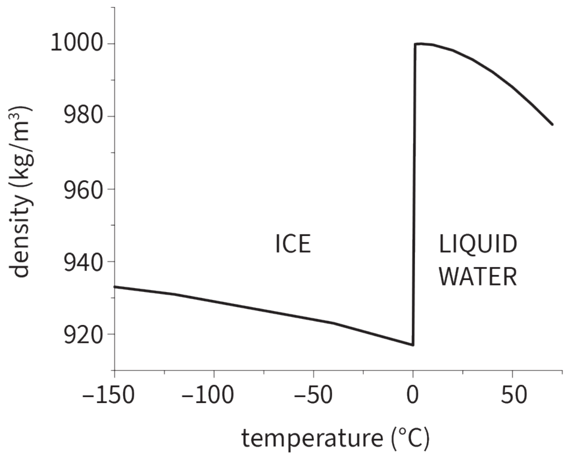


Figure 3.8. Abnormal behavior of water density. Unlike other liquids, where density is higher in the solid state, the density of ice is lower than that of water. Maximum density is observed for water at 4 °C. Data for liquid water sourced from Lemmon et al., while data for solid water was sourced from Kohlrausch (see References section).

liquid phase). Thus, in solid water, the H-bonds are very stable and hold the crystal together, but at the same time they are responsible for the unusual low density of ice compared to liquid water.

Figure 3.8 illustrates the abnormal behavior of water at certain temperatures (called the temperature anomaly of water). Similar to most other substances, water expands when heated. But it does not expand within the range between 0 °C and 4 °C. When ice melts to liquid water (0 °C at normal pressure), the regular network of H-bonds is disrupted, part (but not all) of the open crystalline structure collapses, some molecules break loose and move about the empty spaces. In this way, the overall volume of water decreases (by about 9%) and density increases. As the temperature increases from 0 °C, more ice crystals collapse and the density of water molecules reaches its maximum value at 4 °C. Above this temperature, enough H-bonds have been broken that the liquid starts behaving “normally” (similar to other liquids), becoming less dense with increasing temperature due to greater molecular motion.

When we consider the reverse process, freezing, as the temperature decreases, more and more hydrogen bonds form so that water molecules are arranged in a hexagonal network. Consequently, the volume increases and the density decreases. As solid ice cools further, like most solids, it contracts, but the density of ice at any temperature is less than the density of water which is why ice floats on water.

2.3. Other properties of water

In the body temperature range of homeothermic organisms (35 °C – 41 °C), about 50% of the of the maximal possible number of the water H-bonds are broken, water is less viscous (it flows easier), there are no more crystalline structures, dimers are abundant and water molecules become more chemically reactive.

At temperatures higher than 100 °C, water is in a gaseous (vapor) state. Due to very intense thermal agitation, almost all H-bonds are broken. The majority of water molecules are free, however some dimers can still exist. Compared to other similar chemical compounds (hydrogenated compounds of elements in the same group of the periodic table as oxygen: H_2Te, H_2Se, or H_2S), water has abnormal melting and boiling points.

Extrapolating the properties of these substances to water would predict a boiling point below 0 °C. Water would therefore be a gas under normal temperature and pressure conditions, making it incompatible to life as we know it. There are deviations between predicted and existing values also in relation to other physico-chemical properties of water. The main reasons for these differences are that the water molecule is a dipole and has the ability to form H-bonds. Thus, we can list other special properties of water:

- strong cohesion forces;
- high surface tension;
- high specific heat;
- high heat of vaporization;
- liquid state at room temperature;
- high thermal conductibility;
- high dielectric constant;
- water is a good polar solvent;
- hydrophobic exclusion.

We define cohesion as the attraction between molecules that are alike. Water molecules experience **strong cohesion forces**, due to the H-bonds. This property of water is of physiological importance. The two pleurae coating the thoracic cavity are separated by a thin layer of liquid. The visceral pleura covers the lungs and the parietal pleura covers the inner surface of the thoracic cavity. During inspiration, the volume of the thoracic cavity increases and the lungs expand following the movement of the thoracic wall, due to the strong molecular cohesion of the pleural liquid that causes the simultaneous movement of both pleurae.

With the exception of mercury, water has the highest **surface tension coefficient** of all liquids (for a detailed description, see the last section of this chapter). Because of H-bonds, the cohesion forces between the water molecules forming the superficial layer of water are stronger than the attractive forces between water and air molecules (adhesive forces) and the superficial layer of water behaves like a stretched elastic membrane which

tends to minimize the surface of water-air contact.

Unlike many other substances, water has a **high specific heat capacity**[3], meaning it has to absorb a lot of heat in order to increase its temperature. An increase in temperature would be equivalent to an enhancement of thermal agitation movement of molecules, and, in order for this to happen, some H-bonds need to be broken first. Thus, a large amount of the energy transferred as heat is used for breaking H-bonds. Also, during cooling, water retains more heat than other substances, because the slowing down of the molecules allows new H-bonds to form and, in this process, energy is released. Otherwise put, it takes more time for water to cool down. Due to the high specific heat capacity, the temperature of water can be maintained constant for relatively large energetic changes, and this is of great importance for the thermoregulation of the body.

High heat of vaporization[4] is another property derived from the ability of water to form H-bonds, as water needs a lot of energy to break all the H-bonds so it can evaporate. Actually, this is the reason for which water is in a liquid state at room temperature. Also, water's high heat of vaporization allows us to briefly touch a wetted finger to a hot object without harm.

Water's **thermal conductivity** is several times higher than for most other liquids. From this point of view, water can be seen as a thermal buffer. The thermal inertia caused by the high water content, together with water's high specific heat capacity are important factors for maintaining a constant body temperature, in spite of environmental temperature changes. Local hyperthermias (produced by exothermal catabolic processes) are avoided because of the rapid heat diffusion mediated by water. In addition, the high conductivity of water is the reason for which blood loses a great amount of heat at the body surface, thus allowing the exchange of energy (transferred as heat) between the organism and the environment.

As the water molecule is a permanent dipole that can change its orientation in an electric field, water also has a **high dielectric constant** (high ability to store electrical energy in an electric field), which favors the dissociation of electrolytes. In other words, this is the reason for which table salt and other ionic compounds easily dissolve in water.

Water is a **good solvent for polar substances and substances that can form H-bonds**. Also, in order for molecules of a certain substance to be dissolved and thus leave their solid structure, they must interact more easily with water than between themselves.

3 The *specific heat capacity* is the amount of heat required to change the temperature of a unit mass of substance by 1 °C.

4 *Heat of vaporization* is the amount of heat needed to change the unit mass of substance from liquid to gas and the reverse.

3. AQUEOUS SOLUTIONS AND SYSTEMS

3.1. Water as a solvent

A homogeneous liquid mixture consisting of a single phase is called a *solution*. The component present in the largest amount is the solvent, and any other components are the solutes. We call a solution an *aqueous solution* if the solvent is water.

According to how they behave in their interaction with water, substances can be:

- **hydrophilic** ("water loving"): that easily dissolve in water;
- **hydrophobic** ("water fearing"): do not dissolve well in water;
- **amphiphilic** (also amphipathic): have both hydrophilic and hydrophobic parts.

Electrolytes (ionic salts like NaCl, KCl, etc.) are good examples of **hydrophilic substances**. In contact with water, strong electrolytes completely dissociate into negative ions (anions) and positive ions (cations). The electric fields around ions disturb the orientation of water dipoles; a positive ion attracts water's negative pole (oxygen atom) and a negative ion, water's positive pole (hydrogen atoms). The process is called ion solvation (hydration) and leads to a spherical arrangement of water molecules around the ions, called a hydration shell. The interaction between the ions and the water molecule is a typical example of an ion-dipole interaction. As Figure 3.9 shows, the presence of an ion organizes the water molecules

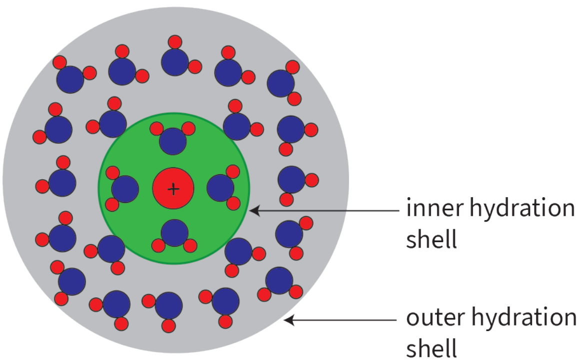


Figure 3.9. Hydration of a cation. A cation is shown in the center. Water molecules arrange themselves in two hydration shells. The inner shell (innermost circle) contains the molecules that directly interact with the ion through ion-dipole interactions. The outer hydration shell (outer circle) contains water molecules that interact with the molecules of the inner shell and to each other through hydrogen bonds.

into an inner (or primary) hydration shell, composed of highly ordered water molecules in the ion's close vicinity, and an outer (or secondary) hydration shell, composed of less ordered water molecules connected via H-bonds to the inner shell. Hydration water has different properties compared to normal water: higher density, different freezing and boiling points, etc.

Generally, the dissolution process in water of a salt such as NaCl can be written as:

$$NaCl_{(s)} \xrightarrow{H_2O_{(l)}} Na^+_{(aq)} + Cl^-_{(aq)} \quad (3.1)$$

where (s) denotes the solid state, (l) the liquid state and (aq) that the ions are dissolved in the resulting aqueous solution.

Other examples of hydrophilic substances are the ones exposing polar groups like –OH or $–NH_2$, that can form H-bonds with water (e.g. alcohols, some amino acids, sugars, etc.). The biological activity of various macromolecules, proteins for example, depends not only on their structure but also on their spatial configuration. This is determined by the H-bonds formed between the protein and the water molecules in its close vicinity or between different parts of the protein. Breaking the H-bonds (for example, by coagulation) leads to irreversible protein denaturation and subsequent loss of function. As stated previously, the stability of the DNA double helix is also given by H-bonds (N–H···N and N–H···O) between the two DNA strands.

Molecules with the ability of forming H-bonds and permanent dipoles can easily integrate into the network of water molecules without producing significant structural changes.

Non-polar molecules interact with water only through weak van der Waals interactions such as Debye forces. They are called **hydrophobic molecules**. The system water-hydrophobic molecule will adopt a configuration with minimum free energy, so a maximum number of bonds will form for this reason. In this context, because water cannot interact via H-bonds with hydrophobic molecules, there will be a higher number of H-bonds between water molecules around the hydrophobic molecules and a crystalline cage-like structure is formed, called a *clathrate*. The hydrophobic molecule is trapped inside the water clathrate. The water molecules making up the clathrate "cage" have reduced mobility, making the arrangement somewhat similar to ice.

The increase in the number of bonds is also achieved by pushing the hydrophobic molecules close to each other, as a result of water's tendency to bond more easily with other water molecules

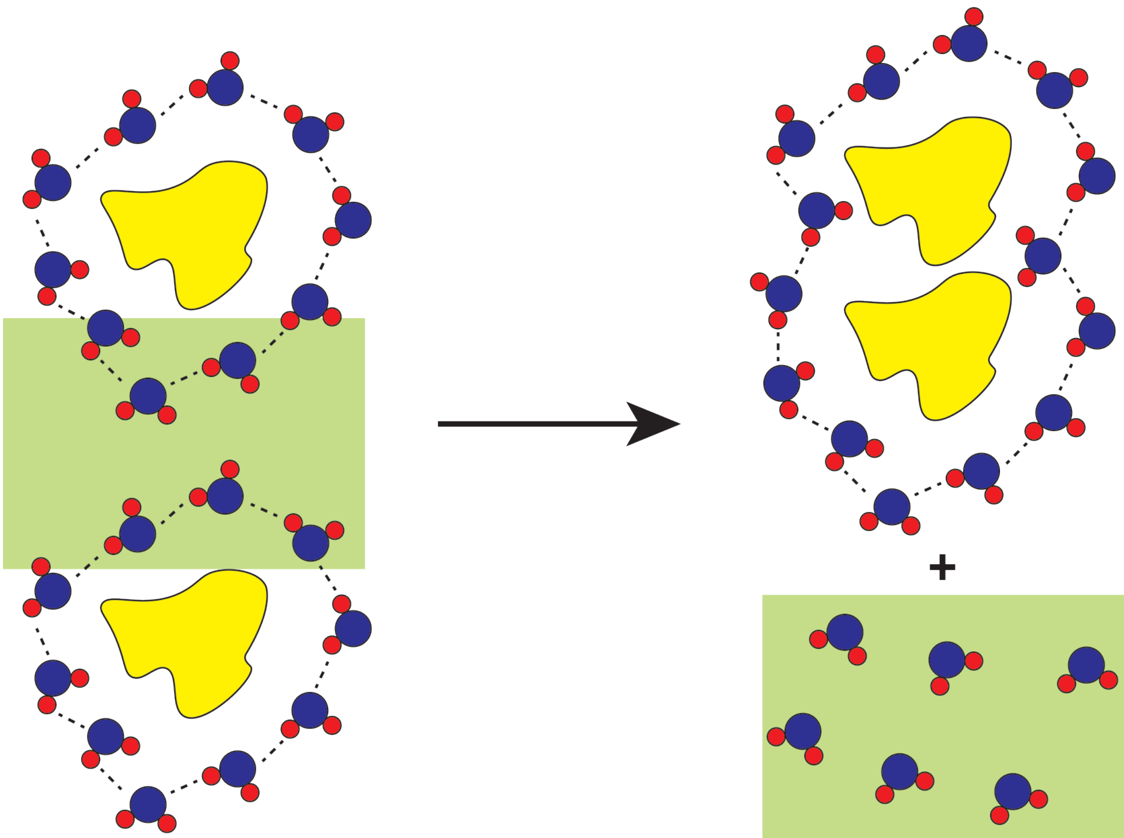


Figure 3.10. Hydrophobic exclusion. Left panel: water molecules form clathrate-like cages around hydrophobic molecules (yellow blobs). It is entropically favorable for the system to evolve (right panel) towards a state where the hydrophobic molecules come close together, as this releases some of the water molecules that were forming the cages (green box) from these structures.

than with the less polar ones. In this way, water surrounds the non-polar solute without sacrificing much of the H-bonds. We can say that the presence of a hydrophobic substance stabilizes the water's H-bonding network which can be pictured as an elastic net compressing the hydrophobic molecules into aggregates that do not mix with water. The process is called **hydrophobic exclusion** or the **hydrophobic effect** (Figure 3.10).

Amphiphilic molecules, like proteins and nucleic acids, contain polar as well as nonpolar groups. The hydrophilic parts will easily bond with other polar molecules or groups through electrostatic interactions, whereas hydrophobic groups are excluded and forced to interact among themselves via van der Waals forces.

3.2. The auto-dissociation of water. pH

The water molecule is remarkably stable. However, there is a small proportion of water molecules that will, under normal conditions, split into two ions: H^+ (hydrogen ion, a proton) and HO^- (hydroxide ion). We call this process the **auto-dissociation** of water (or **auto-ionization**).

The proton is not stable by itself in solution. It will quickly join another water molecule to form a hydronium ion (H_3O^+). In the end, one hydroxide ion and one hydronium ion are formed from two molecules of water:

$$H_2O \leftrightarrows H^+ + HO^- \tag{3.2}$$

$$H^+ + H_2O \leftrightarrows H_3O^+ \tag{3.3}$$

$$2H_2O \leftrightarrows H_3O^+ + HO^- \tag{3.4}$$

In pure water, a water molecule can only accept a proton if another molecule loses a proton. Thus, for every hydronium ion formed, a hydroxide ion also forms, which means that, in pure water, the total number of hydronium ions equals the number of hydroxide ions.

Experiments have shown that, in any aqueous solution, the concentration of hydronium ions multiplied by the concentration of hydroxide ions always equal a very small number (10^{-14} at 25 °C), called the auto-ionization constant of water, K_w. The significance of this constant is of great importance because it means that no matter what is dissolved in water, the product of the hydronium and hydroxide concentrations will always equal 10^{-14} at 25 °C. Therefore, if the concentration of one ion goes up, the concentration of the other must go down. **In pure water, at 25 °C**, their concentrations are equal: **$[H_3O^+] = [HO^-] = 10^{-7}$ mol/L.**

The relative concentrations of hydronium and hydroxide ions determine whether a solution is acidic, basic, or neutral. We call a solution **neutral** if the concentration of hydronium ions equals that of hydroxide ions, as in pure water.

A solution is **acidic** if substances that can act as proton donors (acids) are dissolved. This will increase the concentration of H_3O^+ ions. An example is the dissolution of HCl shown in equation (3.5). As the product of H_3O^+ and HO^- concentrations is fixed, this will also lower the concentration of hydronium ions.

$$HCl_{(g)} + H_2O_{(l)} \longrightarrow H_3O^+_{(aq)} + Cl^-_{(aq)} \tag{3.5}$$

If, instead, the solution contains more HO^- ions than H_3O^+ ions, we call the solution **basic**. An example of a base dissolving in water is shown below:

$$NaOH_{(s)} \xrightarrow{H_2O_{(l)}} Na^+_{(aq)} + HO^-_{(aq)} \tag{3.6}$$

We express the **acidity** of a solution numerically through a quantity called the **pH**. pH is defined as the negative logarithm of the hydronium ion molar concentration:

$$pH = -\log [H_3O^+] \tag{3.7}$$

where log is the logarithm in base 10, and the square brackets denote the concentration of the hydronium ions in solution (at equilibrium).

According to their pH value, aqueous solutions can be:

- **acidic**, if **pH < 7**;
- **neutral**, if **pH = 7**;
- **basic**, if **pH > 7**.

In other words, at pH < 7 the concentration in hydronium ions is higher, whereas at pH > 7, the concentration in hydroxide ions is higher. At pH = 7, the concentrations of H_3O^+ and HO^- are equal, as in pure water.

In the human body, the average pH is approximately 7.4. In biological systems, there are solutes called *buffers* with the ability to maintain the value of pH in a narrow range of values (pH doesn't change too much), a crucial condition for an optimal enzymatic activity. You will learn more about these buffer systems when studying biochemistry and physiology.

4. WATER IN BIOLOGICAL SYSTEMS

In biological systems, water's most important properties are the polar solvability and

hydrophobic exclusion. For example, when immersed into water, lipids self-assemble in membrane-like structures because of hydrophobic exclusion. Also, in an aqueous environment, globular proteins spontaneously adopt a conformation in which the hydrophilic groups are exposed to water and the interactions between hydrophobic regions and water are avoided. In this way, macromolecules organize their own structure as well as the arrangement of the water molecules around them (as hydration and clathrate water). Such a configuration of water molecules is called **bound water**. Studies have shown that inside and outside the cells, water is distributed differently as: free water, partially bound water, and bound water.

A large amount of the water contained by the human body is bound water and, compared to free water (non-bound), it has specific physical properties: it hardly evaporates, it freezes at temperatures well below 0 °C, it does not act as a solvent and it is not involved in osmosis. The behavior of water contained within living organisms is still not completely understood. The state of bound water is induced by the presence of a large number of molecular, macromolecular and ionic species that organize the water molecules through weak electrical dipole-like interactions (between water and various parts of the organic molecule). A high amount of the intracellular water content is in the bound state and it has an important role in cellular processes (excitation, contraction, division, secretion, etc.).

5. INTERFACIAL PHENOMENA

An *interface* is the surface separating two different phases which are in direct contact, such as: an insoluble solid and a liquid, two immiscible liquids, a liquid and a gas, a solid and a gas, etc.

There are two types of intermolecular forces present at the interface level. *Cohesion forces* are attractions between molecules that are alike (e.g. attraction between water molecules), while *adhesive forces* are attractions between unlike molecules (e.g. attraction between water molecules and silica oxide in the glass).

As a consequence of cohesion forces, a surface tension force appears in liquids that are in contact with a gas or with a solid. If the cohesion forces are stronger than the adhesion forces, the surface tension force will resist external deformation of the liquid's surface – essentially, the surface of the liquid will behave in some ways like an elastic membrane.

Surface tension can easily be visualized by performing a simple experiment – fill a cup with water and then take a dry pin or a paperclip and

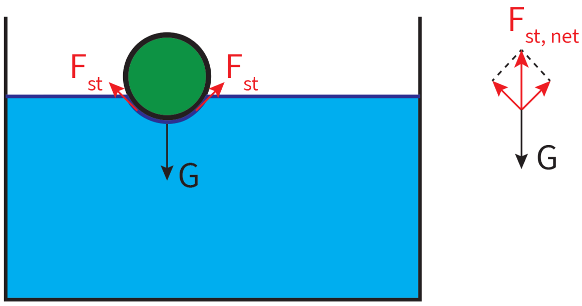


Figure 3.11. A metal pin (of circular cross-section, shown in green), is gently placed on the surface of water (left panel). As long as the pin is light enough, the gravitational force (G) will be balanced by the surface tension forces (F_{st}) which oppose the deformation of the liquid's surface. The right panel shows the net surface tension force balanced against the object's weight. If the balance of forces is disturbed, for example, by pushing down on the pin, then the pin will fall to the bottom of the glass. Note that this experiment only works for light objects; for heavy objects, $G > F_{st}$ and the object will break the water's surface. After the surface of the water is broken, the object's density will dictate whether it floats on the water or drops to the bottom.

gently place it on the surface of the water (Figure 3.11). The light metal object will not, by itself, break the surface of the water as long as the surface is not disturbed. If you gently push down on the metal object, though, it will drop to the bottom of the glass in short order.

One consequence of surface tension is the tendency of liquids to minimize their surface area. For example, small quantities of liquid in a gas take a spherical shape (the minimum surface area for a certain volume).

Note that similar forces appear when two immiscible liquids are in contact. These are called *interfacial tension forces*. While technically named differently, the effect is the same – the interfacial tension acts to minimize the surface area between two immiscible fluids.

The surface tension coefficient, σ, (sometimes called only surface tension[5]) can be measured and has a value for a certain liquid in contact with a certain gas (e.g. air) which also depends on temperature. It is defined as:

- the force (F) required to increase the perimeter of a liquid (l) by one unit:

$$\sigma = \frac{dF}{dl} \tag{3.8}$$

- the work (W) required to increase the surface area (S) of the superficial layer by one unit:

$$\sigma = \frac{dW}{dS} \tag{3.9}$$

5 T or γ are also commonly used as a symbol.

Another consequence of surface tension is *capillary action*, the movement of liquids in narrow tubes or through the pores of a loose material, as a result of cohesive and adhesive forces. Capillary action is very important in many biological processes (e.g. the ascension of water through the xylem vessels in plants, vascular accidents like gas embolism).

Various substances added to the liquid can influence the cohesion forces between the molecules of liquid, thereby altering surface/interfacial tension. Substances with the ability to lower surface/interfacial tension are called **surf**ace **act**ive **a**ge**nts**, or **surfactants**. According to Traube's Law, the surface activity of an amphiphilic molecule increases with the number of hydrophobic components of that molecule. The most effective surfactants are amphiphilic molecules (e.g. detergents, soaps, bile salts, pulmonary surfactant).

In the breathing process, the alveolar capillary wall has a crucial role, as at this level the gas exchange between blood and air takes place. In the lungs, the bronchi branch into bronchioles which, in turn, terminate in alveolar sacs. Each alveolar sac contains a collection of pulmonary alveoli, with an average number of around 700 million and a total surface of 70 – 90 m^2 (by comparison, the total area of the skin is only 2 m^2). The alveoli are the site of the most important contact with atmospheric air. Their surface varies during the respiratory cycle by around 7 m^2. The alveoli are surrounded by a dense capillary network. The alveolar and the capillary epithelia form an extremely thin wall (the thickness is around 0.2 µm). The oxygen and carbon dioxide molecules can cross the alveolar capillary wall by diffusion, down their concentration gradient.

The inner wall of an alveolus is covered by a thin (0.5 µm) layer of liquid which is maintained due to the water vapors present in both inspired and expired air. During the inspiration process, the alveoli inflate due to a higher pressure inside the alveoli relative to their surroundings. This difference in pressure (Δp) depends on the surface tension at the boundary between liquid and air and the radius of the alveolus (r). If we consider the alveoli to be approximately spherical, we can write the relation between these terms according to Laplace's law for a sphere:

$$\Delta \mathrm{p} = \frac{2\sigma}{r} \tag{3.10}$$

If the liquid layer coating the inner surface of the alveoli would be only water, the pressure in an alveolus would be around 12 – 24 Torr.[6] Under this pressure, the air would flow from the small alveoli to the larger ones. Actually, the difference in pressure is smaller, only a few Torr, because the thin layer of liquid also contains pulmonary surfactant, having a phospholipid as main component. The pulmonary surfactant reduces the surface tension and, thereby, the pressure, so there are no significant differences in Δp during the respiratory cycle. Different distribution of surfactant in different sized alveoli leads to a similar pressure in all alveoli, regardless of their size. In this way, the flow of air from smaller alveoli (where, in the absence of surfactant, the pressure would be higher) into the larger ones is prevented. An insufficient amount of pulmonary surfactant or its defective distribution in the alveoli during inspiration can lead to serious respiratory accidents.

6 1 Torr ≈ 1 mmHg.

REFERENCES

Atkins, P. W., De Paula, J., & Keeler, J. (2017). *Atkins' Physical Chemistry*. London: Oxford University Press.

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Flowers, P., Theopold, K., Langley, R., & Robinson, W. R. (2019). *Chemistry 2e*. Retrieved from https://openstax.org/books/chemistry-2e/pages/1-introduction

Glaser, R. (2012). *Biophysics: An Introduction*. Heidelberg: Springer.

Hansen, T. C. (2021). The everlasting hunt for new ice phases. *Nature Communications, 12*(1), 3161. doi:10.1038/s41467-021-23403-6

Kim, M. W., Weon, B. M., & Je, J. H. (2023). Spherical alveolar shapes in live mouse lungs. *Sci Rep, 13*(1), 5319. doi:10.1038/s41598-023-32254-8

Kohlrausch, F. (1996). *Praktische Physik, 3 Bde., Bd.1: Zum Gebrauch für Unterricht, Forschung und Technik* (V. Kose & S. Wagner Eds.). Stuttgart: B.G. Teubner Verlag.

Kontogeorgis, G. M., Holster, A., Kottaki, N., Tsochantaris, E., Topsøe, F., Poulsen, J., . . . Kronholm, J. (2022). Water structure, properties and some applications – A review. *Chemical Thermodynamics and Thermal Analysis, 6*, 100053. doi:https://doi.org/10.1016/j.ctta.2022.100053

Lemmon, E. W., Bell, I. H., Huber, M. L., & McLinden, M. O. Thermophysical Properties of Fluid Systems. In P. J. Linstrom & W. G. Mallard (Eds.), *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*.

Nelson, D. L., Cox, M. M., & Hoskins, A. A. (2021). *Lehninger Principles of Biochemistry Eighth Edition*. New York: MacMillan Learning.

CHAPTER 4

DISPERSION SYSTEMS

Prerequisite knowledge

- Molecules, ions, chemical bonds
- Second law of thermodynamics
- Pressure

1. GENERAL CONCEPTS AND CLASSIFICATION

As described in a previous chapter, water makes up the majority of our body and is indispensable to life. In a typical cell, the cytoplasm is made up mostly of water (~70 – 80%), but it also contains ions, small molecules and macromolecules that are distributed, or *dispersed* inside the water, with water acting as a *dispersion medium*. We call a *dispersion system* or simply a *dispersion* such a mixture where one or several components are distributed in a continuous phase of another component.

Depending on their degree of uniformity, dispersions can be:

- **homogeneous**, if their composition is uniform throughout the system (only one phase exists);
- **heterogeneous**, if their composition is non-uniform throughout the system (two or more phases exist.

Depending on the size of the particles, we classify dispersion systems as:

- **molecular solutions** (also called true solutions or simply solutions) are dispersions in which the size of the dispersed particles is smaller than 1 nm (10^{-9} m). They are fully homogeneous and their components cannot be separated by filtration. Example: table salt dissolved in water;
- **colloids** are heterogenous dispersions where the size of the dispersed particles is 1 nm – 1 µm (10^{-9} m – 10^{-6} m). Macromolecules such as proteins are examples of particles that form colloids. Like solutions, colloids cannot be separated by filtration. Unlike solutions, they scatter a ray of light that is shone into them (the Tyndall effect). Example: milk;
- **suspensions** are heterogeneous dispersions in which the size of the dispersed particles is higher than 1 µm. They can be separated by filtration and settle out if they are left undisturbed (the dispersed particles fall to the bottom of the recipient containing the suspension). Unlike a solution or a colloid, particles in a suspension are visible either with the naked eye or with the use of a simple optical microscope. As in a colloid, the Tyndall effect is visible in suspensions. Example: grains of sand in water.

In the human body, some fluids are at the same time, a solution, a colloid and a suspension. For example, blood is a solution for dissolved ions and gas molecules, a colloid for the macromolecules dispersed inside it and also a suspension for the blood cells.

Colloids can be further classified depending on the state of aggregation of the dispersion medium and the dispersed phase as shown in Table 4.1.

Colloids are also classified depending on the affinity of the phases for each other. Colloids in which the phases have affinity for each other have a high stability and are called *lyophilic*. Examples are gelatin, starch in water, pudding, etc. When the phases do not have affinity for each other, the colloid is called lyophobic. *Lyophobic* colloids are thermodynamically unstable and can only be prepared with high expenditure of energy (for example, by mechanical agitation). Due to their instability, they will aggregate over time. An example of a lyophobic colloid is a metal sol in water (colloidal silver, etc.).

Table 4.1. Classification of colloids.

		Dispersed phase		
		Gas	**Liquid**	**Solid**
Dispersion medium	**Gas**	–	liquid aerosol	solid aerosol
	Liquid	foam	emulsion	sol
	Solid	solid foam	gel	solid sol

2. SOLUTIONS AND THEIR PROPERTIES

2.1. Solutions. The dissolution process

A *solution* is a homogenous mixture, obtained by dissolving and then dispersing one or more chemical species[1] in a dispersion medium called the

1 The term *chemical species* refers to atoms, molecules or ions of the same type.

solvent. The chemical species that are dissolved in the solvent are called *solutes*.

Solutions can have any aggregation state (solid, liquid, gas). Generally, the component which is in the highest quantity in the solution is defined as the solvent. There are, however, some exceptions:

- If only one component in a liquid solution is a liquid, the respective species is the solvent;
- If water is one of the components of a liquid solution, it is generally taken as the solvent (e.g. a mixture of 70 parts ethanol and 30 parts water is called a 70% alcohol solution).

In order for a solute to dissolve in the solvent, it must have some affinity for the solvent. Generally, the dissolution process consists of the following successive steps:

- Solute – solute bonds are broken. This is an endothermic process (ΔH > 0). For example, when dissolving NaCl in water, ionic bonds between the Na^+ and Cl^- ions are broken.
- Solvent – solvent bonds are broken. This is also an endothermic process (ΔH > 0). For example, when substances are dissolved in water, hydrogen bonds between water molecules are broken.
- Solute – solvent bonds are formed. This is an exothermic process (ΔH < 0).

Overall, the net energy balance (sum of enthalpies) of the 3 processes above is calculated, giving the *enthalpy of dissolution* (also called *heat of solution*). In order to establish whether the process is spontaneous, the Gibbs free energy has to be calculated, according to equation (2.15) in the chapter on Thermodynamics. Remember that, for a spontaneous process, ΔG < 0. Thus, the variation of the entropy also has to be considered, not only that of the enthalpy. Generally, dissolution is entropically favored, and can compensate for a slightly positive ΔH. For example, NaCl dissolution is endothermic (ΔH > 0), but the entropy increase during the dissolution process makes it spontaneous (TΔS > ΔH, and, thus, ΔG < 0).

2.2. Concentration

An intensive state parameter that is used to characterize solutions is the *concentration*, which represents the amount of solute (or solutes) dissolved in the solution. As you will see below, there are many methods of expressing concentration, mainly chosen depending on the field of application or the way in which the solution was prepared. As a medical doctor, when prescribing a drug treatment, it is important to always keep in mind what kind of concentration you are working with.

The molar concentration is by far the most common way of expressing concentration:

$$c_M = \frac{\upsilon_{solute}\ (moles)}{V_{solution}\ (liters)} \tag{4.1}$$

where ν_{solute} represents the number of moles of solute dissolved and $V_{solution}$ is the volume of solution expressed in liters.

Thus, the molar concentration represents the number of moles of solvent dissolved in one liter of solution, and is measured in mol/L, also represented as M (molar). For example, a 2 M solution of sucrose in water contains 2 moles of sucrose in a liter of solution.

Other means of expressing concentration are also quite common. For example, percent concentrations are used in many cases. Particular attention has to be given in the case of percent concentrations, as there are different types of these, using either weight (w) or volume (v). These are generally not equivalent to each other:

$$c_{\%}(w/w) = \frac{m_{solute}(g)}{m_{solution}(g)} \cdot 100 \tag{4.2}$$

$$c_{\%}(w/v) = \frac{m_{solute}(g)}{V_{solution}(mL)} \cdot 100 \tag{4.3}$$

$$c_{\%}(v/v) = \frac{V_{solute}(L)}{V_{solution}(L)} \cdot 100 \tag{4.4}$$

Two examples can illustrate the different types of percent concentration:

- the alcohol content of beverages is given as a percent (v/v) concentration. A beer that has a 5% alcohol concentration has 5 mL of ethanol per 100 mL of liquid;
- saline solution is a 0.9% (w/v) solution of NaCl in water. This means there are 0.9 g of NaCl dissolved in 100 mL of solution. This solution is particularly important in medicine because it has a similar osmolarity to blood plasma (osmolarity is another method of expressing concentration which will be defined later in the chapter).

Some examples of other methods of expressing concentration:

- the molal concentration is the number of moles of solute in 1 kg of solvent:

$$c_m = \frac{\upsilon_{solute}\ (moles)}{m_{solvent}\ (kg)} \tag{4.5}$$

- the normal concentration is the number of equivalents of solute dissolved in one liter of solution:

$$c_E = \frac{\upsilon_{equivalents}}{V_{solution}\ (liters)} \tag{4.6}$$

- the molar fraction is the ratio between the number of moles of a component of the solution (either solute or solvent) and the total number of moles of the chemical species in a solution. For

gases, this is equivalent to the partial pressure:

$$X_i = \frac{\upsilon_{component}\,(moles)}{\upsilon_{solution}\,(moles)} \quad (4.7)$$

2.3. Ideal solutions. Solubility

An *ideal solution* is a solution in which the dissolution enthalpy is zero. This is a concept similar to that of an ideal gas. In ideal solutions, interactions between molecules of different components are of the same strength as interactions between molecules of the same component. A solution is closer to ideal when it is more diluted. Most of the simple equations that we will describe further apply to ideal solutions – the more a substance deviates from ideal, the more complex its behavior.

The *solubility* of a chemical species in a certain solvent is defined as the maximum amount of the respective species that can be dissolved. Solubility is affected by the chemical nature of the solute and solvent, the temperature, pressure, and by any other eventual solutes that are dissolved in the solvent. For most solid solutes, solubility in water increases with temperature. A solution that contains an amount of solute equal to the solubility is called a *saturated solution*. If more solute is added to a saturated solution, it will not be able to dissolve and instead remains as a separate phase (for example, if the solute is a solid, it precipitates).

2.4. Solubility of gases in liquid. Medical applications

The solubility of gases in liquid decreases with the increase of temperature. This is because, at higher temperatures, the kinetic energy of gas molecules increases, allowing them to more easily "escape" the solution.

The solubility of gases increases with the increase of their partial pressure. We define partial pressure (Figure 4.1) as the pressure that the molecules of a particular gas in a container would exert if this gas would be alone in the container, and not in a mixture.

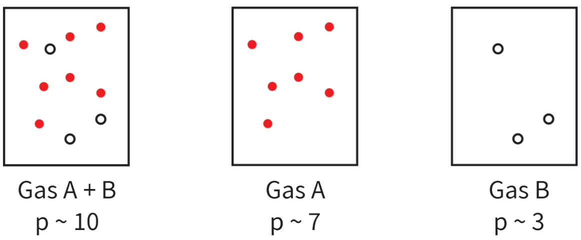


Figure 4.1. A container holds a mixture of molecules of gas A and gas B (A – full circles, B – empty circles) with a ratio of 7 molecules A for each 3 molecules of B. At a given volume, the pressure is directly proportional to the number of particles. If the total pressure in the container is ~10, the partial pressure of A in the mixture is ~7 and the partial pressure of B is ~3.

In a mixture of gases, the sum of the partial pressures of each component is the total pressure of the gas mixture in the container. A simple real-life example of calculating partial pressure is calculating the partial pressure of O_2 and N_2 in the atmosphere. At a normal atmospheric pressure (1 atm) and a proportion of 21% O_2 and 78% N_2 in the air, partial pressures can easily be calculated as p_{O2} = 0.21 atm and p_{N2} = 0.78 atm. The rest of 0.01 atm is given by argon, carbon dioxide and other gases. If the pressures are given in mmHg instead of atmospheres, considering that 1 atm = 760 mmHg, p_{O2} = 160 mmHg and p_{N2} = 593 mmHg. Note that the composition of air in the pulmonary alveoli is different from that of atmospheric air, hence the partial pressures are also different.

Out of the atmospheric gases, carbon dioxide (CO_2) has the highest solubility in water and biological fluids. An excessive amount of dissolved CO_2 in the blood is called *hypercapnia*. This is generally the result of insufficient ventilation (hypoventilation). This can appear, for example, in patients suffering from advanced stages of chronic obstructive pulmonary disease (COPD). Hypercapnia results in a decrease of blood pH (acidosis), due to the formation of HCO_3^- in the blood stream following dissolution of CO_2. Moderately high partial pressures of CO_2 (p_{CO2}) in the alveoli cause shortness of breath (dyspnea). At high p_{CO2} in the alveoli (80 – 100 mmHg), patients suffer from lethargy and can become semicomatose. Partial pressures p_{CO2} of 120 – 150 mmHg result in anesthesia and death.

Lack of oxygen in tissues and cells is termed *hypoxia*. Hypoxia can result either from a lack of available oxygen in the atmosphere (for example, at high altitudes), or by a variety of conditions (pulmonary disease, anemia, etc.). Symptoms are depressed mental activity and reduced muscle capacity. In severe cases, hypoxia leads to death.

An excess of oxygen in the tissues and cells (*hyperoxia*) is also problematic. This can appear when oxygen is breathed in at higher partial pressures than normal. Symptoms can appear at the level of the central nervous system, lungs and eyes. The main effects are believed to occur as a result of the formation of reactive oxygen species (ROS) in the body, that are highly reactive and can damage biologically important molecules (DNA, proteins, lipids).

The dynamics of nitrogen (N_2) solubility also deserve particular consideration, especially in the context of diving. The increase of the partial pressure of N_2 in the alveoli causes an increase in the N_2 concentration in the blood stream and leads to *nitrogen narcosis*, which causes symptoms similar to the first stages of general anesthesia. A decrease in the activity of the central nervous

system is caused by the decrease of neuronal excitability due to the dissolution of N_2 in the neuronal membrane. Nitrogen narcosis can appear in divers that go to low depths while breathing in air, due to the increase of the partial pressure of N_2 in the air mixture. To prevent nitrogen narcosis, divers that go to depths lower than ~30 m have to breathe in mixtures that replace N_2 either partly or totally with helium (He), which has a lower solubility in the lipid membrane.

Another problem that divers can encounter is decompression sickness, which can appear if they return quickly to the surface after diving at low depths. This occurs because, during diving at low depths (high pressures), a high amount of gases (especially N_2) dissolve in the tissues, due to the increased partial pressure of these gases. If the return to low pressures is fast, the solubility of N_2 in the tissues decreases rapidly, and bubbles of gas can appear in body fluids, either intra– or extracellularly, which can cause serious damage. Bubbles of gas can block small blood vessels, leading to ischemia and tissue death. In a few cases, decompression sickness can be fatal. In order to prevent it, divers have to return to the surface slowly, spending sufficient time at certain depths on their way to the surface in order to allow their dissolved N_2 to be slowly released.

2.5. Colligative properties of solutions

Many of the properties of a solution clearly depend on exactly *what* substance you have dissolved in the respective solution. For example, a 0.01 M solution of HCl is acidic (pH = 2) and will give the specific chemical reactions of an acid. A 0.01 M solution of KOH is basic (pH = 12) and will give the specific chemical reactions of a base.

However, there are some properties of a solution where what exactly is dissolved in the solution does not really matter, but what is important is only how much substance is dissolved. We call these *colligative properties*.

Examples of colligative properties are:

- a decrease in the vapor pressure of a solution that contains a volatile solvent and a non-volatile solute compared to the vapor pressure of the solvent alone (Raoult's law);
- a decrease of the freezing point of a solution of a non-volatile solute compared to that of the pure solvent;
- an increase of the boiling point of a solution of a non-volatile solute compared to that of the pure solvent;
- the appearance of an osmotic pressure, which will be described later in this chapter.

The effects of concentration on the freezing point (T_f) and the boiling point (T_b) of a solutionare described by the following equations:

$$\Delta T_f = i \cdot k_{cr} \cdot c_m \qquad (4.8)$$

$$\Delta T_b = i \cdot k_{eb} \cdot c_m \qquad (4.9)$$

where i is a number called the *van't Hoff factor* (detailed below), k_{cr} is the *cryoscopic constant* of the solvent, k_{eb} is the *ebullioscopic constant* of the solvent and c_m is the molal concentration of the solution.

The **van't Hoff factor (i)** represents, in most cases (for a general form, see the section on osmosis), the number of particles that each solute particle produces in the solution. This is easily calculated for solutes that fully dissociate in solution. Let's consider a few examples:

- glucose is a carbohydrate that is easily dissolved in water, but does not dissociate. For each mole of solid glucose added to water, exactly one mole of glucose exists dissolved in the solution. Thus, for glucose, i = 1;
- NaCl is an ionic salt that fully dissociates in water (strong electrolyte). When we add NaCl to water, the ionic bonds holding the NaCl crystals together are broken, and individual ions of Na^+ and Cl^- are formed, which are then hydrated. Consequently, for each mole of NaCl dissolved in water, 2 moles of particles are formed: 1 mole of Na^+ and one mole of Cl^-. Thus, for NaCl, i = 2;
- Na_2SO_4 also fully dissociates in water. When we add Na_2SO_4 to water, the ionic bonds holding the Na_2SO_4 crystal are broken, but the covalent bonds holding together the sulfate ion are not. For each mole of Na_2SO_4 dissolved in water, 3 moles of particles are formed: 2 moles of Na^+ and 1 mole of SO_4^{2-}. Thus, for Na_2SO_4, i = 3.

The way in which substances dissociate in water and, thus, **the value of the van't Hoff factor is extremely important also in osmosis**. Please keep this in mind for the section on osmosis at the end of the chapter.

Analyzing equations (4.8) and (4.9) and the subsequent discussion on the van't Hoff factor, we can draw the following conclusion: for the same concentration of solute dissolved, the effect on the boiling/freezing point of the solution is higher if the value of i is higher. For example, the freezing point of a 1 molal solution of glucose is –1.86 °C. For a 1 molal solution of NaCl, the effect is double: the freezing point is –3.72 °C.

2.6. Electrical properties of solutions

The substances that easily dissolve in water by dissociation into positive and negative ions are called *electrolytes*. *Strong electrolytes* are substances

such as NaCl or NaOH that fully dissociate when added in water. *Weak electrolytes* are substances that only partially dissociate when added in water. For example, acetic acid is a weak electrolyte that dissociates in water as follows:

$$CH_3COOH + H_2O \rightleftarrows CH_3COO^- + H_3O^+ \quad (4.10)$$

Each molecule of CH_3COOH dissociates into one acetate ion (CH_3COO^-) and one proton (H^+, which is not stable by itself in solution, and forms a hydronium ion together with a molecule of water). We can define for every weak electrolyte the **degree of dissociation**, noted as α, which can be calculated as:

$$\alpha = \frac{amount\ of\ substance\ dissociated}{amount\ of\ substance\ initially\ dissolved\ in\ solution} \quad (4.11)$$

For example, for a 0.1 M solution of acetic acid, $\alpha = 0.01$, or otherwise said only about 1% of acetic acid is actually dissociated into acetate ions, while the rest remains undissociated in solution.

Solutions of electrolytes differ significantly from pure water:

- the electrical resistance of electrolyte solutions is much weaker than that of pure water;
- an electrical potential difference ΔV appears between solutions of different concentrations.

Ionic strength is a measure of the concentration of ions in a particular solution. It can be calculated as:

$$I = \frac{1}{2}\sum C_i z_i^2 \quad (4.12)$$

where C_i is the concentration of each ion in the solution and z_i the electrical charge of the ion. As can be evident, multivalent ions contribute strongly to ionic strength. Ionic strength is very important when solutions that deviate from ideality are described.

Another quantity used to describe non-ideal (real) solutions is the activity. Activity represents the "effective" concentration of the solution, which has to be used when particles of solute are abundant enough to have non-ideal interactions with each other and with the molecules of solvent. Activity is defined as:

$$a = f \cdot \frac{C}{C^0} \quad (4.13)$$

where C is the concentration of the solution, C^0 is the standard state for concentration (1 molal or 1 molar) and *f* is the activity coefficient. For ideal solutions, the activity coefficient $f = 1$.

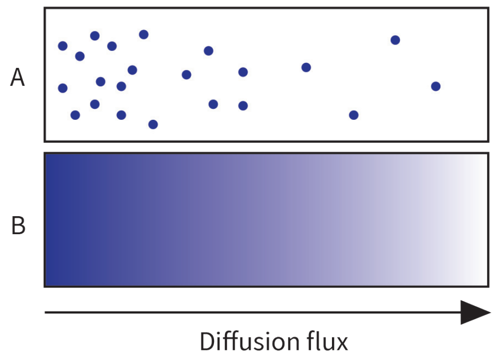


Figure 4.2. Diffusion in a simple system. A, microscopic view of a simple system (container + a few molecules of solute). Molecules of solvent are not represented, for simplicity. A concentration gradient leads to a diffusion flux. B, macroscopic view of the same type of system. Higher concentrations are shown as darker shades of blue.

3. DIFFUSION

3.1. Definition. Laws of Fick

We define *diffusion* as the transport of solute within the volume of solution, as a result of the existence of a concentration gradient[2] (Figure 4.2). As previously discussed in the Biological thermodynamics chapter, the existence of any gradient in a thermodynamic system will lead to the appearance of a flux that acts to diminish the respective gradient. Diffusion is, thus, a spontaneous process, which occurs in the direction in which the Gibbs free energy of the system decreases, according to the second law of thermodynamics.

Diffusion is described quantitatively by equations called the **laws of Fick**. The **first law of Fick** can be used for a system in a steady state, where the concentration gradient is constant over time. Such systems are very common in living organisms, and, as discussed previously, a living cell is a good example of a system in a steady state. The first law of Fick can be written as:

$$J = -D\frac{dC}{dx} \quad (4.14)$$

where *J* is the diffusion flux, *D* is the diffusion coefficient and *dC/dx* is the concentration gradient.

The sign minus shows that the diffusion flux acts to make the gradient disappear. However, as discussed before, in the steady state the gradient is held constant by expenditure of energy from a different source. Thus, the diffusion flux also stays constant. The diffusion coefficient is a quantity that describes how easily particles of solute are free to move (diffuse) in solution. For a simplified, spherical particle, the diffusion coefficient can be calculated as:

2 For charged solutes we have to consider also the existence of electrical potential gradients.

$$D = \frac{kT}{6\pi\eta r} \tag{4.15}$$

where k is Boltzmann's constant (this constant has a deep physical meaning and it is related to the Boltzmann microscopic definition of entropy; it happens that the ideal gas constant, R, is the product of k and Avogadro's number N_A), T is the temperature in Kelvin, η is the viscosity of the solution and r is the radius of the particle.

When the concentration gradient changes as a result of diffusion occurring, the **second law of Fick** must be used. This is written as:

$$\frac{dC}{dt} = -D\frac{d^2C}{dx^2} \tag{4.16}$$

where dC/dt is the variation of concentration over time, D is the diffusion coefficient and d^2C/dx^2 is the variation of the concentration gradient in space.

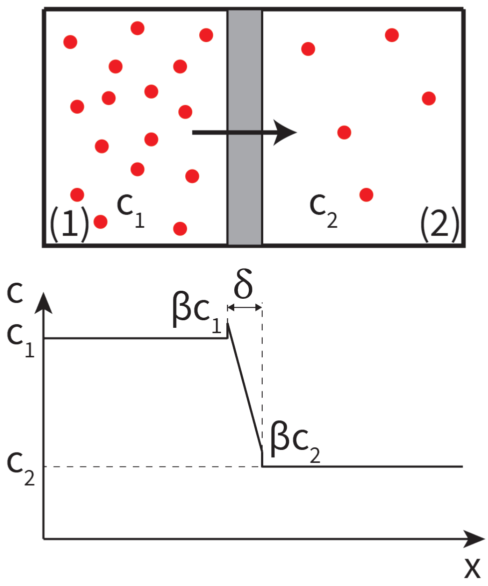


Figure 4.3. Diffusion over a biological membrane. Two compartments (1) and (2) contain solute (with β > 1) of different concentrations $c_1 > c_2$. δ = thickness of the membrane, β = partition coefficient of the solute.

3.2. Diffusion through membranes

A membrane is defined as a thin layer that separates two compartments. Biological membranes generally separate media of different composition (for example, the cytoplasm from the extracellular medium). Thus, concentration gradients are usually present between the two compartments that the membrane separates.

The most important criterion used to classify membranes is what chemical species can pass through them, or otherwise said, their selectivity. Accordingly, membranes are classified as:

- **permeable** (equally permeable if all components of the solution can diffuse as easily, or unequally permeable, if different components have different permeabilities);
- **selectively permeable** (if only some components can pass through the membrane);
- **semi-permeable** (the solvent can pass through the membrane);
- **non-reciprocal** (solutes can pass through the membrane, but only in one direction);
- **impermeable** (no component can pass through the membrane).

Note that a membrane can belong to several of the above categories. Biological membranes are considered both semi-permeable membranes (water can easily pass through them) and selectively permeable (only certain solutes are allowed to pass).

We will treat below only the diffusion of non-electrolytes through membranes, while the diffusion of charged species will be described in the following chapter on Membrane transport.

Let us consider a system in which two compartments (1) and (2) are separated by a biological membrane of thickness δ such as that in Figure 4.3. A solute concentration gradient exists between the two compartments, and the concentration of the solute is higher in compartment (1), so $c_1 > c_2$.

In order for solute to diffuse from compartment (1) to compartment (2), the solute has to be able to dissolve in the lipidic environment of the membrane. Generally, the concentration of the solute in the membrane is not the same as that in solution, because the solubility of the solute in the hydrophobic interior of the membrane is different from the solubility in an aqueous (water-based) solution. We can thus define the **partition coefficient** of the solute, β, as:

$$\beta = \frac{solubility\ of\ the\ solute\ in\ the\ membrane}{solubility\ of\ the\ solute\ in\ the\ solvent} \tag{4.17}$$

We can define a second quantity called the permeability coefficient of the solvent, P as:

$$P = \frac{D\beta}{\delta} \tag{4.18}$$

where D is the diffusion coefficient of the solute, β the permeability coefficient and δ the thickness of the membrane.

At the interface between the membrane and each compartment, the concentrations of the solute will thus be βc_1 and βc_2, respectively. Thus, at the level of the membrane, the concentration gradient is $(\beta c_1 - \beta c_2)/\delta$. If the system is in a steady state, we can consider the gradient as constant in time and we can apply Fick's first law. We thus find that the diffusion flux is proportional to the difference in concentration between the two

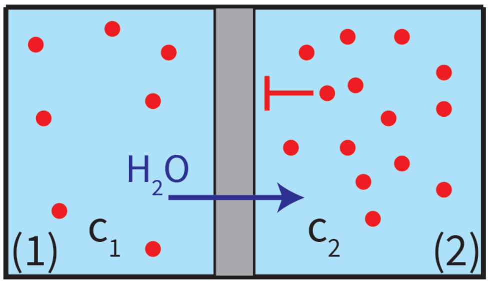


Figure 4.4. Osmosis. Two compartments (1) and (2) contain solute of different concentrations $c_2 > c_1$ and are separated by a semi-permeable membrane that is impermeable to the solute. Thus, the solute cannot diffuse, but a flux of solvent appears, from the compartment that has the most solvent (1), towards the compartment that has the least solvent (2).

compartments:

$$J = -P\Delta C \tag{4.19}$$

In a living cell, both active and passive fluxes of solute exist. Only the latter are attributable to diffusion, and thus described by the above equation. Furthermore, as mentioned before, the above is only valid if the solute has no electrical charge.

4. OSMOSIS

4.1. Definition. Osmotic pressure

Consider two compartments containing solutions of different concentration separated by a semi-permeable membrane that is impermeable to the solute (Figure 4.4). The second law of thermodynamics dictates that the spontaneous direction for the system to evolve into is the equalization of concentrations in the two compartments (in order to maximize entropy). As the solute cannot pass through the membrane, the solvent will move, but in the opposite direction, from where the solute is less concentrated to where the solute is more concentrated.

The transport of solvent through a semi-permeable membrane, from a more diluted solution to a more concentrated solution, is called **osmosis**.

We can thus consider that the molecules of solvent (water) in the left compartment of Figure 4.4 are "pushed" into the rightmost compartment. We could stop this flux of water molecules by applying sufficient pressure from the right compartment in order to "push" back against the flow of water. The pressure that we would need to apply in order to stop the osmotic flow is called *osmotic pressure*.

Osmotic pressure (noted with π) can be calculated differently depending on the type of solute dissolved in the solution. Note that only the concentration of osmotically active particles (particles that cannot pass through the membrane) contributes to the osmotic pressure in the equations below.

For **ideal solutions of non-electrolytes**, osmotic pressure can be calculated using the law of van't Hoff, which is very similar in form to the law of ideal gases. This law can be written as:

$$\pi V = \nu RT \tag{4.20}$$

$$\pi = RTC \tag{4.21}$$

where π is osmotic pressure, V is the volume, ν is the number of moles, T is the temperature, R is the constant of ideal gases and C is the molar concentration of the solution.

If solutions containing different amounts of solute are separated by a semi-permeable membrane, the osmotic flow is determined by the difference in the osmotic pressures of the two solutions:

$$\Delta\pi = RT\Delta C \tag{4.22}$$

As we discussed previously, osmosis is a colligative property of solutions, thus (with an important exception in the case of macromolecules, discussed below), the identity of the dissolved solutes does not matter for the value of the osmotic pressure, only the number of particles dissolved in solution does.

For **ideal solutions of electrolytes**, the number of particles that the solvent dissociates into has to be considered. We already introduced this number as the van't Hoff factor and gave some examples of how it can be calculated for a non-dissociating substance, as well as for a strong electrolyte. If one considers weak electrolytes as well, the van't Hoff factor is defined as:

$$i = 1 + (p - 1)\alpha \tag{4.23}$$

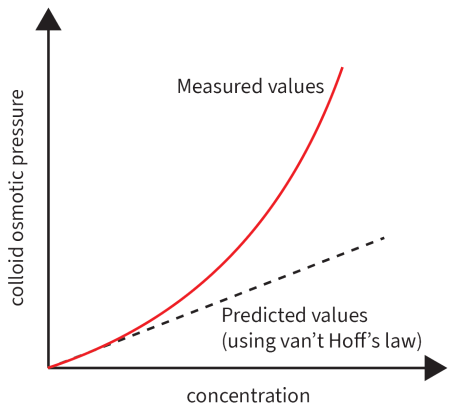


Figure 4.5. Colloid osmotic pressure of a protein. The measured curve deviates from the theoretical linear pressure dependence calculated using van't Hoff's law.

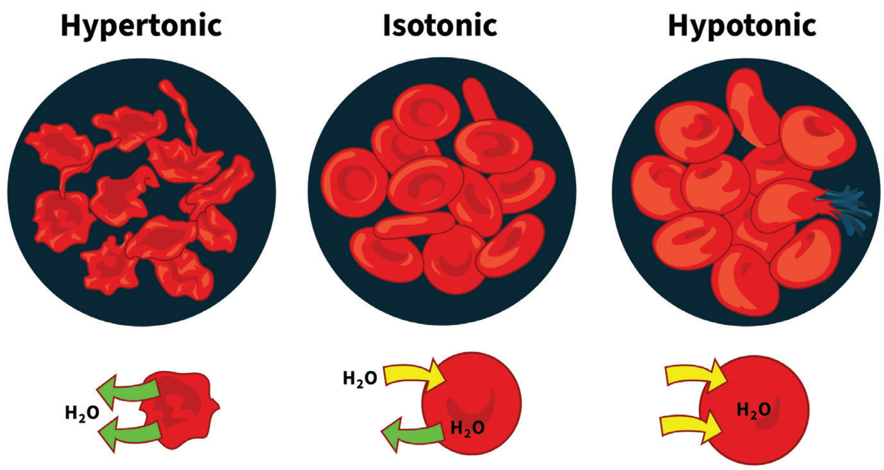


Figure 4.6. Effect of placing red blood cells in solutions of different tonicities.[3]

where p is the number of ions that a solute particle dissociates into when dissolved and α is the degree of dissociation of the solute. Note that for strong electrolytes, $\alpha = 1$, thus $i = p$.

In the case of **colloidal dispersions** (containing **macromolecules** such as proteins), the measured values of the osmotic pressure are higher than the values predicted by van't Hoff's law (Figure 4.5). Thus, the contribution of macromolecules to the osmotic pressure is higher than that of chemical species of a smaller size. The osmotic pressure of macromolecule solutions is termed **colloid osmotic pressure** or **oncotic pressure**. Colloid osmotic pressure is extremely important for capillary exchange, as will be discussed further.

4.2. Osmolarity

As discussed above, in the case of colligative properties such as osmosis, the **number of particles dissolved per unit of volume** is important, rather than the number of molecules that were added per unit volume of solution (the molar concentration). The quantity that expresses that is called the **osmolar concentration**, or **osmolarity**.

We define an **osmole** as the number of osmotically active particles dissolved in solution equal to the number of Avogadro (N_A). The osmolar concentration (osmolarity) represents the number of osmoles (Osm) dissolved in 1 liter of solution:

$$osmolarity\ \left(\frac{Osm}{L}\right) = \frac{no.\ of\ osmoles\ dissolved}{V_{solution}\ (liters)} \quad (4.24)$$

Some examples of calculating osmolarity are given below:

- A sucrose solution of molar concentration 1 mol/L has an osmolar concentration of 1 Osm/L, because the sucrose molecule does not dissociate when dissolved in water, thus 1 mole of sucrose forms 1 osmole when dissolved;
- A potassium chloride (KCl) solution of molar concentration 1 mol/L has an osmolar concentration of 2 Osm/L. This is because each mole of (solid) KCl dissolves into two osmoles when added in water: 1 osmole of K^+ and 1 osmole of Cl^-;
- A calcium chloride ($CaCl_2$) solution of molar concentration 1 mol/L has an osmolar concentration of 3 Osm/L. This is because each mole of (solid) $CaCl_2$ dissolves into three osmoles when added in water: 1 osmole of Ca^{2+} and 2 osmoles of Cl^-.

5. DIFFUSION AND OSMOSIS IN LIVING ORGANISMS

5.1. Plasma osmolarity. Tonicity

Cellular membranes have a high permeability for water, derived from two sources. First, as they are quite small, water molecules can permeate through the lipid bilayer. Second, most cells express in their membrane transmembrane proteins that facilitate the passage of water molecules called *aquaporins*.

The membrane of red blood cells is highly permeable to water. Therefore, red blood cells are very sensitive to changes in the osmolarity of their surrounding medium. In order for red blood cells to survive, they have to be surrounded by a solution of the same osmolarity as their cytoplasm. Thus, the **osmolarity of blood plasma**[4] is highly regulated in the body and normally has a **value of ~300 mOsm/L (0.3 Osm/L)**.

We call a solution of the same osmolarity as

3 Modified from public domain image by LadyofHats available at *https://commons.wikimedia.org/wiki/File:Osmotic_pressure_on_blood_cells_diagram.svg*

4 To be more precise, the normal osmolarity of blood plasma is between 280 and 300 mOsm/L.

another solution an *isotonic* solution. For blood plasma, an isotonic solution is one that has an osmolarity of ~300 mOsm/L. Solutions of lower osmolarities than the reference solution (< 300 mOsm/L in the case of blood plasma) are called **hypotonic** while solutions of higher osmolarities than the reference solution (> 300 mOsm/L in the case of blood plasma) are called **hypertonic**.

What happens if red blood cells are placed in solutions of different osmolarities (Figure 4.6)? Placing red blood cells in an isotonic solution produces no ill effects, as expected. When red blood cells are placed in a hypertonic solution, water will quickly leave the cells due to osmotic flow. As a consequence, the cells shrink and eventually die. When red blood cells are placed in a hypotonic solution, a high influx of water will cause the cells to swell and, eventually, burst.

It should be, then, immediately apparent that, **in order to prevent damage to red blood cells, fluids that are administered intravenously need to be isotonic with blood plasma**. Commonly used isotonic solutions are saline solution (0.9% NaCl in water), 5% glucose solution or Ringer's lactate solution (a solution containing Na^+, Cl^-, lactate, K^+ and Ca^{2+} of osmolarity ~273 mOsm/L).

5.2. Capillary exchange

Capillary exchange represents the exchange of gases, nutrients and metabolites to and from the capillaries and the surrounding interstitial fluid. The capillary wall is composed of a single layer of endothelial cells, separated by spaces called *intercellular clefts*, which serve as passageways for substances to diffuse to and from the capillary. Small polar and nonpolar molecules, such as CO_2 or O_2 can diffuse easily through the lipid bilayer of membranes and thus can pass directly through the membranes of the capillary endothelial cells. Ions and larger molecules such as glucose can only diffuse through the clefts. However, macromolecules such as plasma proteins are generally not able to pass through the clefts, and are thus are the only species that are osmotically active.

The direction in which the fluid exchange occurs across the capillary wall is controlled by four parameters termed Starling forces. These are (Figure 4.7): the capillary pressure, the interstitial fluid pressure, the capillary plasma colloid osmotic pressure and the interstitial fluid colloid osmotic pressure. Out of the four Starling forces, note that the capillary pressure varies across the length of the capillary (dropping as one moves from the arterial end to the venous end), while the other three remain the same along the length of the capillary.

We can group the Starling forces according to their nature. We'll call the blood pressure in the capillary the *hydrostatic pressure* in order to distinguish it from the colloid osmotic pressure. Thus, the **hydrostatic pressure difference** across the capillary wall can be calculated as:

$$\Delta p = p_{capillary} - p_{interstitial\ fluid} \quad (4.25)$$

The **colloid osmotic pressure difference** can be calculated as:

$$\Delta\pi = \pi_{capillary} - \pi_{interstitial\ fluid} \quad (4.26)$$

Simply put, we can say that the hydrostatic pressure difference (Δp) is what drives water and dissolved substances out of the capillary, while the colloid osmotic pressure difference ($\Delta\pi$) is what draws water and dissolved substances into the capillary.

Thus, the net balance of Starling forces is what controls the direction in which substances are transported. This is called the **net filtration pressure** (NFP), which, according to the classical Starling equation, can be calculated as:

$$NFP = \Delta p - \Delta\pi \quad (4.27)$$

It has been observed, though, that the capillary wall does not perfectly exclude proteins. Thus, a *reflection coefficient* (σ) is calculated, describing the ratio between the observed and theoretical values of the colloid osmotic pressure difference:

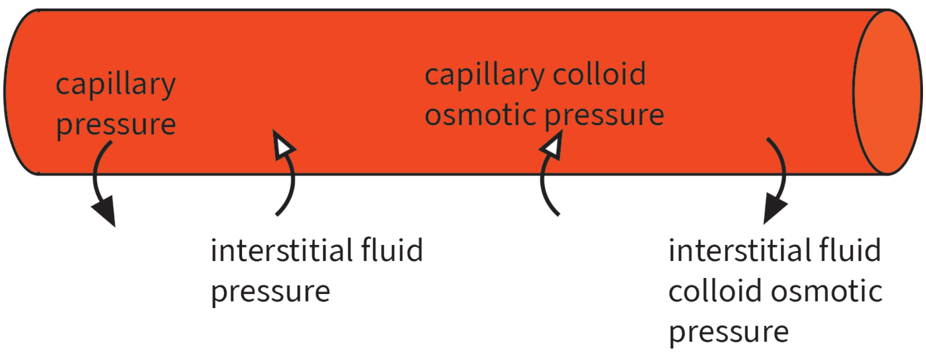


Figure 4.7. Starling forces and the direction in which each force drives transport. Detailed description is provided in the text.

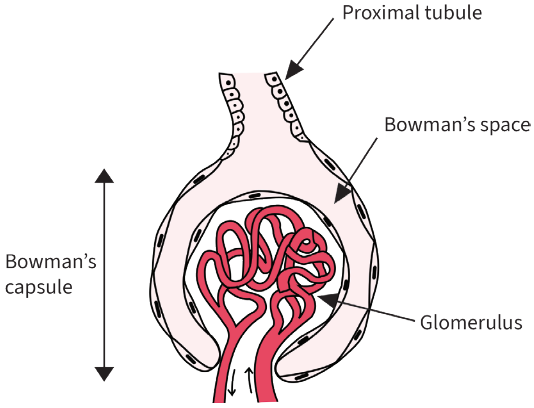


Figure 4.8. Bowman's capsule.[5]

$$\sigma = \frac{\Delta\pi_{observed}}{\Delta\pi_{theoretical}} \quad (4.28)$$

At the extremes, if σ = 0, proteins are fully permeable through the capillary wall and, thus, osmotically inactive, while σ = 1 corresponds to no permeability of the proteins. Taking the reflection coefficient into account, the net filtration pressure is calculated as:

$$NFP = \Delta p - \sigma\Delta\pi \quad (4.29)$$

Two distinct situations appear in different portions of the capillary:

- At the arterial end of the capillary, the NFP is positive ($\Delta p > \sigma\Delta\pi$), therefore water and nutrients leave the capillary. This process is called **ultrafiltration**.
- At the venous end of the capillary, the NFP is negative ($\Delta p < \sigma\Delta\pi$). This is because, as previously, mentioned, the capillary pressure drops significantly from the arterial end to the venous end of the capillary, while Δπ remains constant. Thus, liquid from the interstitial space enters the capillary. This process is called **reabsorption**.

To give some numerical examples[6], Δp can vary from 35 mmHg at the arterial end of the capillary down to 15 mmHg at the venous end, while Δπ = 25 mmHg and σ ≈ 1 for plasma proteins. Thus, using equation (4.29), NFP = 10 mmHg at the arterial end of the capillary and NFP = –10 mmHg at the venous end.

The most important contributors to the value of the capillary colloid osmotic pressure are the plasma proteins, mainly albumins. These cause the colloid osmotic pressure of blood to be higher than that in the interstitial fluid. An abnormally decreased level of plasma proteins, such as that caused by certain kidney diseases in which plasma proteins are passed into urine, can lead to a decrease in the plasma colloid osmotic pressure. This causes increased capillary filtration, leading to edema (accumulation of excess fluid in tissues).

5.3. Renal ultrafiltration. Dialysis

The balance of Starling forces controls exchange of nutrients in the kidney. Kidney ultrafiltration is performed in the portion of the nephron called Bowman's capsule, which contains a tuft of capillaries called the glomerulus (Figure 4.8).

As before, we can calculate the NFP according to equation (4.29), replacing the pressures in the interstitial space with those in the glomerular capsule and considering a reflection coefficient for proteins σ ≈ 1. An NFP value of ~10 mmHg is estimated in humans. Thus, fluid is driven through ultrafiltration from the glomerulus into the nephron. Glomerular filtration is then followed by reabsorption of water and some substances such as amino acids, glucose, Na^+, K^+ which happens in the tubular portion of the nephron.

Severe loss of kidney function is treated by *dialysis*, which has the role of removing toxic products from the blood and maintaining a normal blood composition. This is generally performed by circulating the patient's blood through a machine called a dialyzer. In the dialyzer, blood is separated by a semipermeable membrane from the dialyzing fluid and runs in the opposite direction from the dialyzing fluid. The semipermeable membrane is porous, allowing all constituents of the plasma except for the plasma proteins to diffuse in either direction. By balancing the composition of the dialyzing fluid, toxic waste products such as urea or creatinine can be removed, while preventing loss of nutrients such as glucose. The process of dialysis is lengthy, however, and patients have to undergo the procedure several times per week. As nephrons do not regenerate, the only other treatment for permanent loss of kidney function is renal transplant.

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell 4th Edition: International Student Edition*. New York: Garland Science.

Atkins, P. W., De Paula, J., & Keeler, J. (2017). *Atkins' Physical Chemistry*. London: Oxford University Press.

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară

5 Modified from public domain image from Gray, H. (1918). *Anatomy of the Human Body*. Philadelphia: Lea & Febiger, available at *https://commons.wikimedia.org/wiki/File:Gray1130.svg*

6 According to Boron, W. F., & Boulpaep, E. L. (2017). *Medical Physiology* (3 ed.). Philadelphia: Elsevier.

Carol Davila.
Boron, W. F., & Boulpaep, E. L. (2017). *Medical Physiology* (3 ed.). Philadelphia: Elsevier.
Flowers, P., Theopold, K., Langley, R., & Robinson, W. R. (2019). *Chemistry 2e*. Retrieved from https://openstax.org/books/chemistry-2e/pages/1-introduction
Gray, H. (1918). *Anatomy of the Human Body*. Philadelphia: Lea & Febiger.
Guyton, A. C., & Hall, J. E. (2005). *Textbook of Medical Physiology. Eleventh Edition*. Philadelphia: Elsevier.
Kontogeorgis, G. M., Holster, A., Kottaki, N., Tsochantaris, E., Topsøe, F., Poulsen, J., . . . Kronholm, J. (2022). Water structure, properties and some applications – A review. *Chemical Thermodynamics and Thermal Analysis, 6*, 100053. doi:https://doi.org/10.1016/j.ctta.2022.100053

CHAPTER 5

MEMBRANE TRANSPORT

Prerequisite knowledge

- Second law of thermodynamics. Spontaneous processes
- Gradients and fluxes. Diffusion
- Basic biochemistry of lipids and proteins

1. STRUCTURE OF THE CELL MEMBRANE

1.1. The cell membrane – overview

A clear boundary separates the interior of a cell from the extracellular medium – the *cell membrane*. The cell membrane is a complex structure that fulfils several roles for the cell:

- it serves as a physical barrier that separates the cytosol from the extracellular medium;
- it is selective, allowing only certain chemical species to pass, while blocking the passage of others;
- it is able to respond to external signals and receive information from its environment.

The structural backbone of the cell membrane is given by small lipid molecules, arranged as a double layer. Inserted in the lipid bilayer are large protein molecules, which mediate most of the membrane's functional roles in terms of transport and communication. Generally, the thickness of a membrane is between 6 to 10 nm.

While all cellular membranes have the same general structure (lipid bilayer and membrane proteins), the exact composition of each membrane is different in terms of lipid and protein content and identity, according to the individual characteristics of the cell that it borders. Furthermore, intracellular membranes also exist, bordering organelles such as the mitochondria or nucleus, each with its own particular lipid and protein composition.

1.2. The lipid bilayer

If the number of molecules is taken as the main criterion, lipid molecules make up the majority of the cellular membrane. From a chemical standpoint, three classes of lipids are present in cell membranes: *phospholipids, glycolipids* and *cholesterol*.

The main class of lipid making up cell membranes are the phospholipids, which are built on a backbone of either glycerol (phosphoglycerides) or sphingosine (sphingomyelin).

Phosphoglycerides are esters of glycerol in which the first two hydroxyl groups of glycerol are esterified to carboxyl groups of two fatty acids, while the third hydroxyl group is esterified to phosphoric acid. The phosphate group of the phosphoglyceride can either remain as such to form phosphatidate, which is a minor component of the membrane, or it can be esterified with an alcohol group of the following compounds, forming the corresponding phospholipid:

- choline → phosphatidylcholine, see Figure 5.1;
- serine → phosphatidylserine;
- glycerol → phosphatidylglycerol;
- inositol → phosphatidylinositol;
- ethanolamine → phosphatidylethanolamine.

If the backbone of the lipid molecule is made up of a sphingoid base, such as sphingosine, the lipid is called a sphingolipid. Figure 5.2 shows the structure of a simple sphingolipid, sphingomyelin.

Because of their chemical structure, phospholipids are amphiphilic molecules. Thus, the side of the molecule containing the phosphate group carries a negative charge and is, therefore, hydrophilic. This is termed the hydrophilic head of the lipid (Figure 5.3). The two hydrocarbon

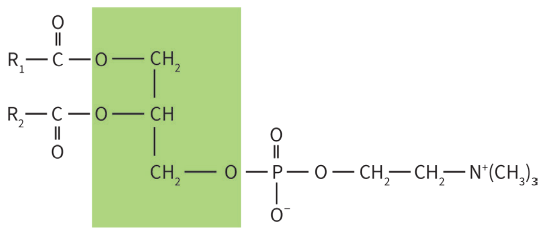


Figure 5.1. Phosphatidylcholine. The glycerol backbone of the lipid is highlighted. R_1 and R_2 are hydrocarbon chains.

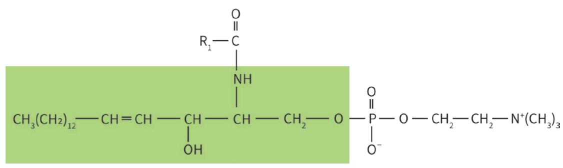


Figure 5.2. Sphingomyelin. The sphingosine backbone of the lipid is highlighted. R_1 is a hydrocarbon chain.

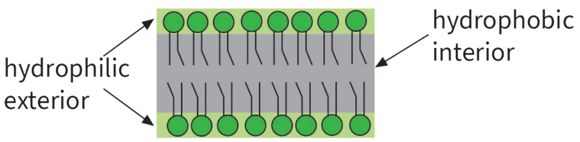


Figure 5.3. A lipid bilayer. The inside of the bilayer contains the hydrophobic lipid tails. The outside of the bilayer is hydrophilic, because of the presence of the polar lipid heads.

chains of the fatty acids of the phospholipid run approximately parallel to each other. Due to their chemical nature, they are hydrophobic and are termed the hydrophobic tails of the lipid (Figure 5.3).

The formation of the lipid bilayer is a consequence of the amphiphilic nature of the lipid molecules – due to **hydrophobic exclusion** (see Figure 3.10), these molecules arrange themselves in a way that tries to minimize as much as possible the contact between the hydrophobic lipid tails and water. This organization is spontaneous (occurs without expenditure of energy). Depending on the shape of the amphiphilic molecule, either micelles (Figure 5.4A) or bilayers (Figure 5.4B) are formed; the cylindrical form of the phospholipid molecules favors the formation of bilayers.

Hydrophobic exclusion is, thus, the main contributor to the formation and stability of the bilayer. Additionally, smaller contributions to the bilayer's stability come from weak intermolecular interactions (van der Waals forces) between the lipid tails as well as from electrostatic interactions between the lipid heads and water molecules in the cytoplasm and extracellular medium.

As mentioned previously, phospholipids are not the only lipid molecules present in the membrane. Glycolipids are lipids derived from sphingosine in which sugars are attached to the hydroxyl group of the sphingosine backbone (instead of phosphocholine). In the membrane, these are found in the outer leaflet of the bilayer, with their sugar groups oriented on the extracellular side.

Cholesterol is the third major lipid component of the membrane, and its main role is that of

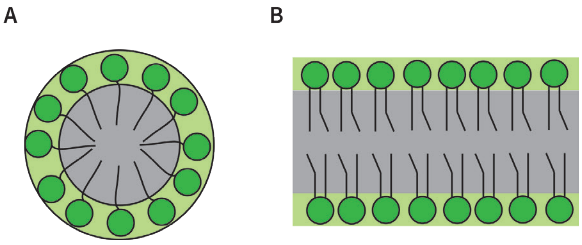


Figure 5.4. Self-organization of amphiphilic molecules due to the hydrophobic effect. A, Micelle formed by a typical amphiphilic molecule (with a single hydrophobic tail). B, Lipid bilayer.

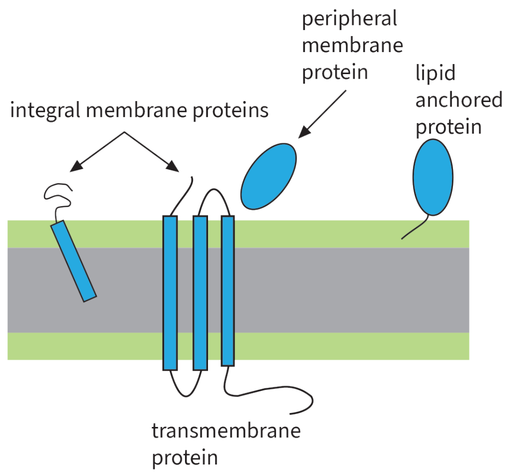


Figure 5.5. Types of membrane proteins – cartoon representation. Integral membrane proteins enter the membrane. Out of these, transmembrane proteins completely traverse the membrane and have both extra- and intracellular portion. Peripheral membrane proteins are weakly attached to the membrane (typically, through temporary interactions with other membrane proteins). Lipid anchored proteins are covalently attached to lipids.

regulating the fluidity of the membrane bilayer. Under most conditions, cholesterol reduces the fluidity of the membrane, by decreasing the mobility of the $-CH_2-$ groups that are close to the polar heads of the phospholipids. However, cholesterol can also increase membrane fluidity, by preventing phospholipid molecules from packing closely together and crystallizing at lower temperatures.

1.3. Membrane proteins

Although far fewer in number compared to the lipids, membrane proteins make up the main component by weight of the cellular membrane. While lipids form a physical permeability barrier, membrane proteins are responsible for most of the functions of the cellular membrane. For this reason, more than 60% of the drugs currently in development target the function of membrane proteins. We will describe in detail some of the roles of membrane proteins later in this chapter.

Membrane proteins (Figure 5.5) can be classified depending on the type of interaction they have with the membrane:

- If they are only weakly attached to the membrane (for example, through van der Waals interactions with other membrane proteins), they are called **peripheral membrane proteins**.
- If at least a portion of the membrane protein enters the lipid bilayer, these are called **integral membrane proteins**. Integral proteins are amphiphilic, with the portions that are inserted inside the membrane containing mainly non-polar amino acid residues, while the intra- and

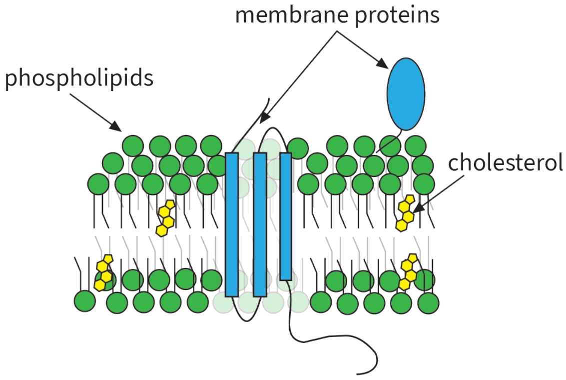


Figure 5.6. Fluid mosaic model of the cell membrane.

extracellular portions contain mainly polar amino acid residues. A special category of integral membrane proteins is represented by **transmembrane proteins**, which span the entirety of the cell membrane. Transmembrane proteins are very hydrophobic and generally aggregate in water, needing detergent for solubilization.

▸ A third class of membrane proteins consists of proteins that covalently attach to the lipids of the membrane. These are called **lipid anchored proteins** or lipid-linked proteins.

1.4. The fluid mosaic model

Biological membranes were found to be extremely fluid structures. Thus, lipid molecules have a high mobility and are able to easily perform lateral movements (lateral diffusion). By comparison, the proteins are less mobile. Transverse movement of lipids (from one side of the membrane to the other) is, however, very slow, and requires specialized membrane proteins to facilitate it.

Overall, the currently accepted model of the cellular membrane is that of the *fluid mosaic* (Figure 5.6), proposed by Singer and Nicholson in 1972. In this model, membranes act as bidimensional solutions of lipids and proteins, with the fluidity regulated by cholesterol molecules. Inside the membrane, some regions contain more cholesterol and sphingolipids. These regions are more rigid than the rest of the bilayer and are called *lipid rafts*.

2. CLASSIFICATION OF TRANSPORT MECHANISMS. MACROTRANSPORT

2.1. Microtransport and macrotransport

Every cell must be able to communicate with its environment and thus needs to have mechanisms in place that allow passage through the cellular membrane. While some chemical species can diffuse through the lipid bilayer (described later), most cannot and require an alternate method of crossing the membrane. This is done through the means of either membrane proteins, or vesicles.

Thus, transport of a larger amount of substance at the same time is mediated by vesicles and is termed macrotransport or bulk transport. Transport of a single molecule or ion at a time is performed by either simple diffusion through the lipid bilayer or by specialized membrane proteins and is termed microtransport.

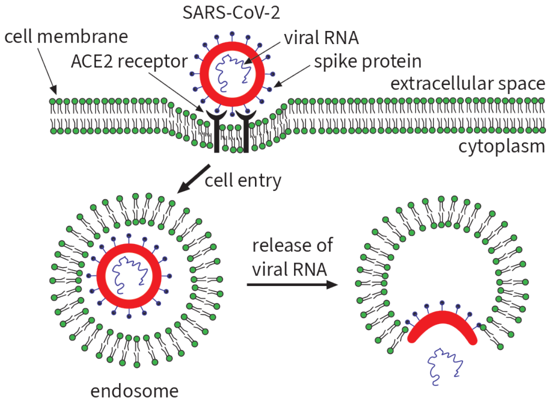


Figure 5.7. Entry of the SARS-Cov-2 virus in the cell is mediated by binding to the ACE2 receptor. Vesicles (endosomes) are formed through receptor-mediated endocytosis, engulfing the viral particle. Viral RNA is then released from endosomes into the cell.

2.2. Classification of macrotransport

Macrotransport can be performed through the following three mechanisms: *endocytosis, exocytosis* and *transcytosis.*

In **endocytosis**, large particles are taken up by the cell. Thus, the membrane fully encloses the particle, after which the membrane buds, forming a vesicle that contains the particle. The detached vesicle is then free to move inside the cell. Depending on the type of particle taken up and the mechanisms involved, endocytosis is further classified as phagocytosis, pinocytosis and receptor-mediated endocytosis.

Phagocytosis is performed by specialized cells of the innate immune system such as macrophages and neutrophils which engulf foreign particles (such as bacteria, parasites or cell fragments) and sequester them in compartments called phagosomes which then fuse to lysosomes that destroy the foreign particles.

Pinocytosis is a non-specific process in which the cell takes up a portion of the extracellular fluid and is used, for instance, in taking up large amounts of nutrients such as amino acids or sugars.

Receptor-mediated endocytosis is a highly specific process in which cells capture certain molecules that bind to specialized receptors.

This is mainly used by the cells for taking up high quantities of molecules that are present at a low concentration in the extracellular fluid. However, receptor-mediated endocytosis can also be hijacked by viruses, which use the process to enter the cell. For example, the SARS-CoV2 virus that causes COVID-19 can enter the cell through the binding of its spike protein to the ACE2 receptor present on the surface of human cells (Figure 5.7).

Exocytosis represents the bulk transport of molecules out of the cell. Vesicles containing the substances to be released fuse to the membrane and release their contents in the extracellular medium. One example of exocytosis can be found in synapses, when neurotransmitter molecules are released in the synaptic cleft (this is described in the following chapter).

Transcytosis is a method of transporting substances across the entire cell. Molecules are engulfed in vesicles at one side of the cell, transported inside the vesicles to the other side of the cell, and then released. Transcytosis happens often in epithelial cells, as these separate two distinct environments.

3. THE ELECTROCHEMICAL POTENTIAL GRADIENT

Movement of ions or molecules over the membrane is strictly regulated by the cell. We classify the microtransport mechanisms that allow chemical species to pass through the membrane without energy expenditure as mechanisms of **passive transport**. Conversely, if the cell needs to expend metabolic energy in order to drive transport, this is called **active transport**.

The difference between passive and active transport can also be easily explained using what we know from thermodynamics – a system will move spontaneously (with no outside energy input) towards thermodynamic equilibrium, thus explaining passive transport. In order to move a thermodynamic system further away from equilibrium, however, energy needs to be spent.

For an uncharged molecule, the direction of passive transport is governed by the concentration gradient of that molecule, which serves as the driving force of the transport process. Thus, passive transport occurs from high to low concentration (or high to low partial pressure in the case of gases). This is also called *downhill transport*.

A more detailed analysis has to be given to the movement of ions, as these interact electrically with their environment and are thus affected by the existence of an imbalance of charges (electrical potential) in the system. Thus, for ions (either large or small), movement through passive transport will be governed by the value of their electrochemical potential gradient.

The electrochemical potential is, in effect, the potential energy that a particular chemical species has in solution, and contains two terms, one dependent on the concentration (chemical potential), and a second that depends on the electrical potential.

The chemical potential (energy) of a solution can be defined as the chemical energy (Gibbs free energy) of 1 mole of substance and is calculated as:

$$\mu = \mu_0 + RT \ln c \tag{5.1}$$

where μ_0 is the standard chemical potential (potential at a concentration of 1 mol/L), R is the constant of ideal gases, T is the temperature (in K) and c is the concentration. While you can find the value of the standard chemical potential of a certain chemical species in databases, this is generally not necessary, as we are usually only concerned with the *variation* of the chemical potential, as defined below in equation (5.4).

The electrical potential energy can be calculated as:

$$W_{el} = zFE \tag{5.2}$$

where z is the electrical charge of the ion, F is Faraday's constant (the electric charge carried by one mole of monovalent ions) and E is the electrical potential of the solution.

Please note that, despite their close names, W_{el} and E are different quantities! W_{el} is, physically speaking, a molar energy (measured in J/mol), while E is the electrical potential (measured in V).

By adding up the two previous equations, we can thus, define the electrochemical potential as:

$$W = \mu_0 + RT \ln c + zFE \tag{5.3}$$

Note that, if a chemical species has no charge, the last term is null (as z = 0) and thus, the electrochemical potential is given just by the chemical

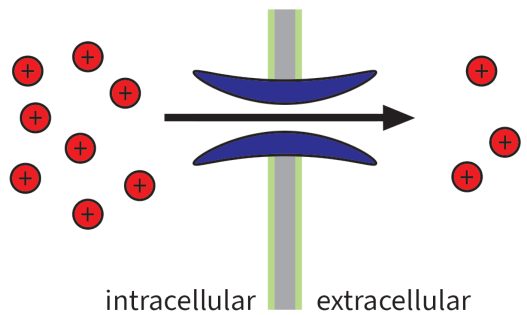


Figure 5.8. The electrochemical potential gradient. Transport of an ion over the membrane can occur only if there is a membrane protein that mediates transport.

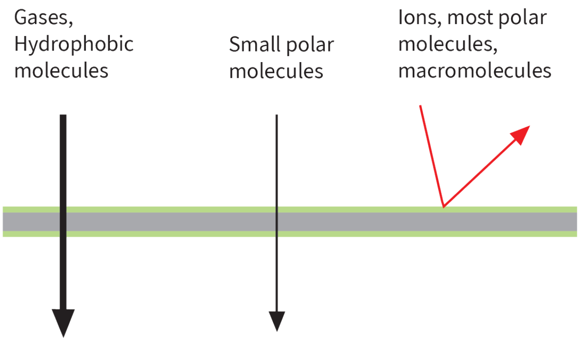


Figure 5.9. Simple diffusion through the lipid bilayer of the membrane. Thinner arrows for polar molecules show lower permeability than that for gases and hydrophobic molecules.

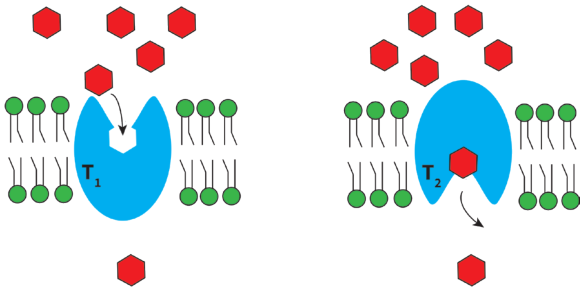


Figure 5.10. Transport through a uniporter. Left - The substrate (hexagon) can bind to its dedicated binding site on the uniporter, which is initially in the T_1 conformation (open to one side). Right - Following binding, the uniporter switches to the T_2 conformation (open to the other side), where the substrate is in lower concentration and can detach from the uniporter.

potential energy term.

In order to calculate the electrochemical potential gradient, one must calculate the difference between the electrochemical potentials of a given chemical species on either side of the membrane (see Figure 5.8). Note that transport will occur only if there is a way for the ion or molecule to actually pass through the membrane (the membrane is not impermeable to that chemical species).

For a system such as that in Figure 5.8, we can calculate the difference in electrochemical potential over the membrane as:

$$\Delta W = W_{ex} - W_{in} = RT \ln \frac{c_{ex}}{c_{in}} + zF(E_{ex} - E_{in}) \quad (5.4)$$

Passive transport will occur from the compartment where the electrochemical potential is higher to where it is lower. Thus:

- if $W_{ex} < W_{in}$, then $\Delta W < 0$ and passive transport will move the ion out of the cell;
- if $W_{ex} > W_{in}$, then $\Delta W > 0$ and passive transport will move the ion into the cell;
- if $W_{ex} = W_{in}$, then $\Delta W = 0$ and there will be no net transport of the ion.

In the latter case ($\Delta W = 0$), the ion has reached equilibrium (no net transport occurs). We can calculate the electrical potential at which this ion is at equilibrium using the **Nernst equation**:

$$E = \frac{RT}{zF} \ln \frac{c_{ex}}{c_{in}} \quad (5.5)$$

where $E = E_{in} - E_{ex}$ is called the **equilibrium potential** of the ion (also called **Nernst potential** or **reversal potential**).

4. PASSIVE TRANSPORT

4.1. Simple diffusion

Due to its hydrophobic interior, the lipid bilayer is virtually impermeable to ions and the majority of polar molecules, while allowing hydrophobic molecules to diffuse through. An exception to this rule is the water molecule, which can permeate the bilayer due to its small size (Figure 5.9).

As discussed previously, diffusion of the permeable species occurs in the direction governed by the concentration gradient, or the pressure gradient if the diffusing molecule is a gas.

4.2. Uniport

In order to facilitate the passive transport over the membrane of chemical species that have low or no permeability in the lipid bilayer, the cell can employ two types of specialized integral membrane proteins: **uniporters** or **channels**.

A **uniporter** belongs to a class of integral membrane proteins called membrane transporters. Uniporters are built in such a way to be able to bind a specific chemical species (also called the **substrate** of the transporter) on one side of the membrane and then, by changing their conformation, move this chemical species to the other side of the membrane.

In a simplified manner, we can draw the uniporter as making a hydrophilic pocket in the membrane (Figure 5.10), where the substrate can bind without being exposed to the highly hydrophobic interior of the membrane, as the substrate binding site contained in this pocket has high affinity for the substrate. After substrate binding, the uniporter suffers a conformational change that closes the access to the side of the membrane where the transporter was bound, and opens access to the opposite side. Subsequently, the substrate detaches from the uniporter, which is free to cycle back.

A characteristic of the uniporters is their high specificity. Thus, a uniporter can only transport its specific substrate, and usually not even

other, highly similar chemical species can be transported.

An example of an essential uniporter is Glucose Transporter 1 (GLUT1), which is expressed almost all tissues in the human body, with particularly high amounts present in the plasma membrane of erythrocytes and that of endothelial cells of the blood-brain barrier. GLUT1 ensures the fast absorption of glucose, at a rate of ~1200 molecules of glucose/s carried across the membrane by each transporter molecule. Homozygous loss of the Glut1 gene is lethal to the embryo, while loss of function mutations in the gene lead to an infantile-onset neurodevelopmental disorder termed the Glut1 deficiency syndrome.

From a kinetic point of view, membrane transporters such as uniporters, symporters and antiporters act as enzymes (Michaelis-Menten kinetics, described in detail in Biochemistry). Note that the last two types facilitate active transport and will be described later in this chapter. Compared to enzymes, where the substrate is changed by the enzyme to an end product, in membrane transport the substrate stays the same, while the transporter changes its conformation. As for enzymes, you can define a maximum transport rate (v_{max}) and a Michaelis constant (K_m), with the general equation defining transport rate being:

$$v = v_{max} \frac{[substrate]}{[substrate] + K_m} \tag{5.6}$$

4.3. Channels

An alternative method of facilitated diffusion through the membrane is performed by **channels**, dedicated membrane proteins that serve essentially as selective "holes" through the membrane, allowing chemical species to flow at a very high transport rate. Most often, channels transport ions, and are, thus, called **ion channels**. However, there are also channels for non-ionic species (e.g. water channels = *aquaporins*). Compared to transporters, ion channels are much faster (more than a thousand times more), but are slightly less selective.

Figure 5.11 shows a general schematic of two types of ion channels. Typically, channels are large proteins assembled from several subunits. Most ion channels contain one or more regions of the protein called **gates** that control whether the passageway through the membrane is open or not. Opening of the gate(s) is regulated by another

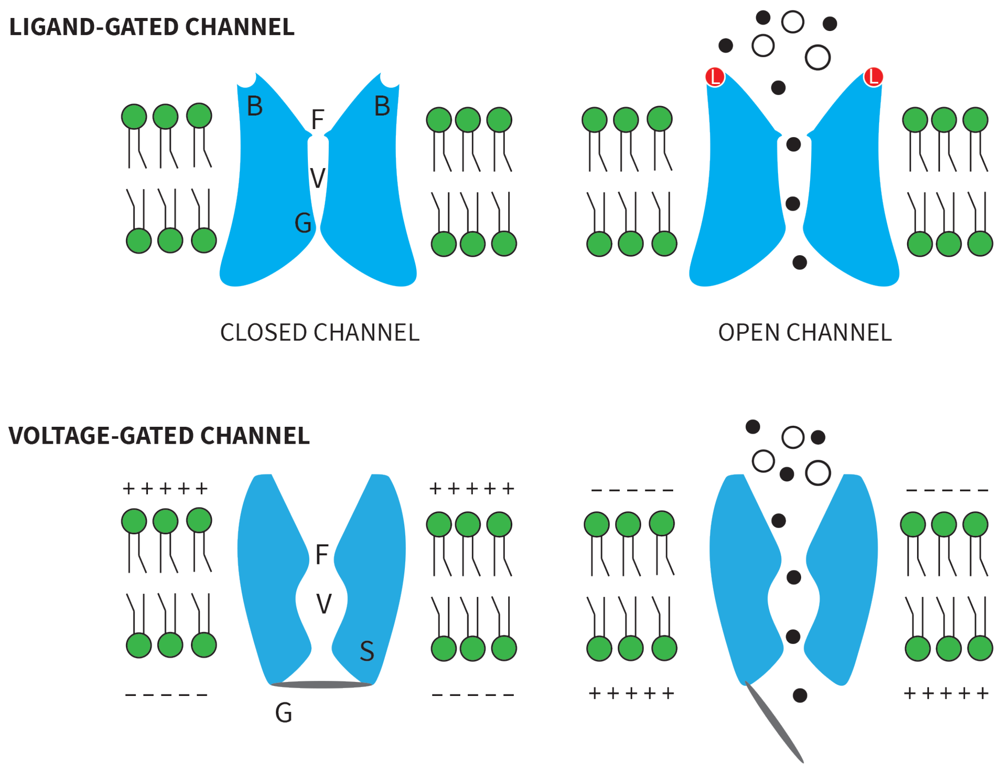


Figure 5.11. Cartoon schematic of two types of ion channels in the closed and open conformation. Top panel: a ligand-gated ion channel. The ligand (L) binds to a specific ligand binding site (B), allowing the channel to open its gate (G). Bottom panel: a voltage-gated ion channel. The sensor region (S) is a group of amino acids sensitive to changes in the membrane potential. When the sensor detects a particular change in the membrane potential (for example, a depolarization of the membrane), it triggers a conformational transition that opens the gate (G). In both panels, the ions normally transported by the channel (black circles) are allowed to pass through the vestibule (V) by the selectivity filter (F), while other ions (white circles) are not.

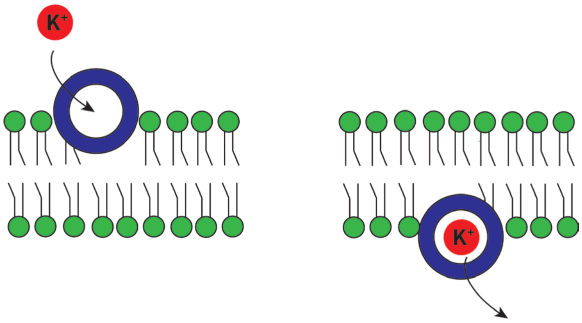


Figure 5.12. Example of an ionophore – valinomycin (black circle). Left – Valinomycin is permeable through the lipid bilayer. K^+ ions can bind to valinomycin, which screens ("hides") their charge from the hydrophobic interior. Right – valinomycin passively diffuses through the lipid bilayer, releasing K^+ on the other side of the membrane.

region of the channel protein called the **sensor**, which can be activated by a specific stimulus that depends on the type of channel involved. To list just a few, the stimulus can be a specific chemical species, termed a ligand (for ligand-gated channels), a change in the membrane potential (for voltage-gated channels), a photon (for light-activated channels in bacteria, algae and fungi), etc. Ions can only enter the main opening of the channel (hydrophilic cavity = the **vestibule**) if they are allowed to pass by the **selectivity filter** region, which blocks undesired ions. The interior of the vestibule is hydrophilic, thus allowing the transit of ions, while bypassing the hydrophobic barrier of the lipid bilayer.

The reason why the transport rate through channels is much faster (~10^6 ions/second) compared to that of transporters (at most ~10^3 molecules/second) is that there is no necessity of always switching conformation in a channel. Once the channel's gate is open by stimulation of the sensor, ions can freely move through the vestibule.

Ion channels are involved in most aspects of a cell's function and are, thus, indispensable. Deficiencies in a specific type of ion channel are termed channelopathies. For example, Na^+ and K^+ channels are involved in regulating the response of excitable cells such as neurons to external stimuli, as will be described the following chapter. Deficiency in certain ion channel genes expressed in the brain can lead to epilepsy.

4.4. Other means of facilitated diffusion

In Gram-negative bacteria that have two membranes (inner and outer), larger structures similar to ion channels exist in the outer membrane, called *pores*. Unlike ion channels, pores allow much larger substrates to pass and have very low selectivity, acting essentially as sieves that block all molecules that are larger in size than the pore opening, while allowing all others through.

Bacteria and other pathogenic microorganisms are also capable of producing molecules called pore-forming proteins (PFPs) that act as toxins that insert into the plasma membrane of the host cells and permeabilize the membrane in order to promote colonization and spread.

Ionophores are another class of substance that can allow passage of ions through the plasma membrane. An example is valinomycin (Figure 5.12), a naturally occurring antibiotic produced by *Streptomyces* bacteria. Valinomycin is a small molecule that has a hydrophobic exterior, but carries an internal hydrophilic pocket capable of selectively binding K^+ ions. Thus, as valinomycin is soluble in the lipid bilayer, it is capable of transporting K^+ (but not Na^+) across the membrane.

5. ACTIVE TRANSPORT

5.1. General principles of active transport

There are many cases when a solute needs to be transported against its electrochemical gradient. As an example, our bodies are built to absorb as much glucose as possible from the nutrients we digest, even when the glucose concentration in the intestinal lumen is lower than that in the cytoplasm of the epithelial cells of the intestinal lining (enterocytes). In order to bring the glucose into the enterocytes from a lower to a higher concentration, the cell needs to expend metabolic energy, and thus facilitate a form of **active transport**. This is also called *uphill transport*.

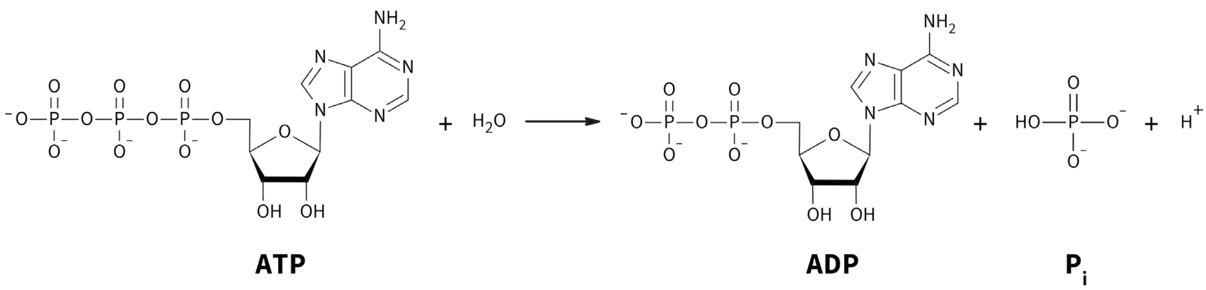


Figure 5.13. Hydrolysis of ATP to ADP and inorganic phosphate releases a high amount of energy ($\Delta G < 0$).

The energy expended to perform active transport can sometimes be spent directly by the protein that facilitates the transport process, for example, by breaking down ATP into ADP and inorganic phosphate. In such a case, we call the respective process **primary active transport**.

In other cases, the energy required for active transport might be provided by a parallel process that occurs spontaneously. For instance, the transport of glucose against its gradient performed in the small intestine mentioned in the first paragraph is performed by transporters termed Sodium/glucose cotransporters (SGLT), which, at the same time, transport Na^+ ions according to their gradient and glucose against its gradient. This is an example of **secondary active transport**.

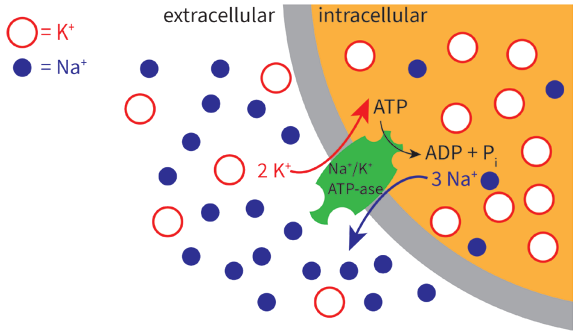


Figure 5.14. The Na^+/K^+ ATP-ase establishes and maintains the Na^+ and K^+ concentration gradients across the plasma membrane. In each cycle of function, the Na^+/K^+ ATP-ase consumes one molecule of ATP in order to move two K^+ ions into the cell (against the K^+ gradient) and 3 Na^+ ions out of the cell (against the Na^+ gradient). Overall, for each cycle a net negative charge appears in the cell (two positive charges are introduced while 3 are removed). In the picture, 3 binding sites for Na^+ and 2 for K^+ are suggested on the ATP-ase cartoon by white semicircles.

5.2. Primary active transport

The main effectors of primary active transport in the human body are membrane proteins that facilitate active transport by directly expending energy stored in the chemical bonds of the ATP molecule, thus breaking down ATP (Figure 5.13) into ADP and P_i (inorganic phosphate). These proteins are termed **ATP-ases** or **pumps**.

The most important ATP-ase is the Na^+/K^+ ATP-ase that is absolutely essential for the correct function of all cells in our body. One molecule of Na^+/K^+ ATP-ase (Figure 5.14) moves, in each transport cycle, 3 Na^+ ions outside the cell, and brings 2 K^+ ions into the cell, at the expense of consuming one molecule of ATP.

Why is the Na^+/K^+ ATP-ase essential? In order to properly function, all our cells rely on the existence of an electrical potential difference across the plasma membrane termed the membrane potential. The membrane potential is generated by the asymmetrical distribution of ions on either side of the plasma membrane, two of the main chemical species involved being Na^+ and K^+ ions. Thus, Na^+ ions are more concentrated in the extracellular medium compared to the cytoplasm, while K^+ is more concentrated in the cytoplasm. These concentration gradients are generated and maintained by the Na^+/K^+ ATP-ase, which drives the transport of both ions against their electrochemical gradients at the expense of consuming

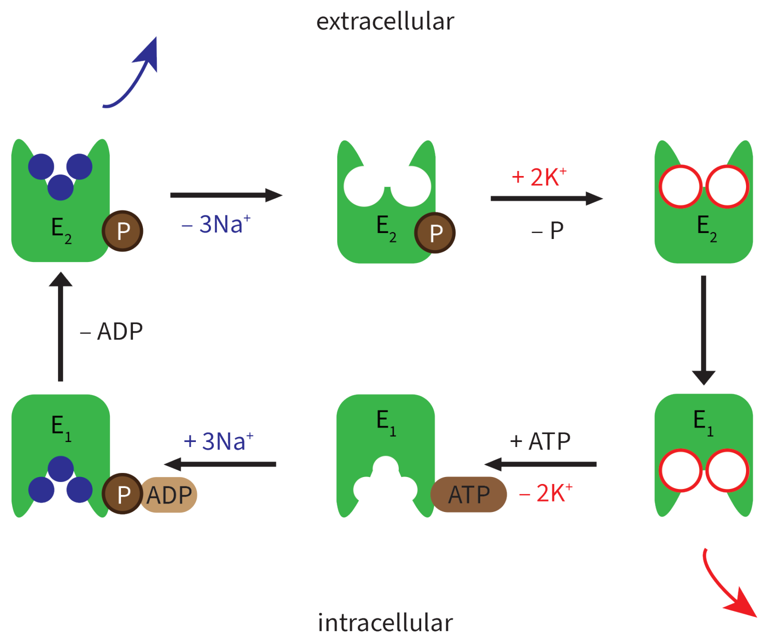


Figure 5.15. A schematic Albers-Post cycle for the Na^+/K^+ ATP-ase. The protein has one of two possible conformations, E_1 (inside-open) or E_2 (outside-open). The complete cycle is described in the text. Na^+ ions = small, filled circles; K^+ ions = large, empty circles; P = inorganic phosphate.

ATP. In brain neurons, the energy expenditure of the Na^+/K^+ ATP-ase accounts for as much as 75% of the total energy requirements of the cell.

Without going into a fully detailed explanation of how the membrane potential is generated, which is reserved for a future chapter, we can easily see that, for each cycle it functions, the Na^+/K^+ ATP-ase creates an imbalance of electrical charge, removing 3 Na^+ ions from the cell for each 2 K^+ ions it brings in, thus generating one net negative charge inside the cell. We can, therefore say that the Na^+/K^+ ATP-ase is an **electrogenic pump**. Note that the opposite of electrogenic would be **electroneutral** = a protein which moves no net charge over a full transport cycle.

The Na^+/K^+ ATP-ase functions according to the Albers-Post cycle (Figure 5.15). In brief, the ATP-ase exists in two main conformations, E_1 (open to the cytoplasm) and E_2 (open to the extracellular space). We will describe the cycle in Figure 5.15 starting from the bottom middle state. In the E_1 conformation that has bound ATP, 3 Na^+ ions can bind to the protein. Hydrolysis of ATP by the E_1 form in the presence of Mg^{2+} ions causes phosphorylation of the pump. Detachment of ADP forces a conformational transition that switches the Na^+-bound protein from the E_1 to the E_2 conformation. In the E_2 conformation, the protein is no longer able to bind Na^+ and releases it, but it is now able to bind 2 K^+ ions. Once K^+ binds and the protein is dephosphorylated, the conformation switches from the E_2 to the E_1 form, K^+ is released inside the cytoplasm and ATP is bound. The protein can now bind Na^+ again, starting a new cycle.

Other essential ATP-ases exist in the human body. Mentioning only a few of them: the K^+/H^+ ATP-ase is responsible for the acid gastric secretion, the Ca^{2+}/H^+ ATP-ase of the sarcoplasmic reticulum (SERCA) has a critical role in muscle contraction as will be described in a future lecture, while the F_1F_o ATP synthase works in reverse compared to the most ATP-ases, using the movement of ions in order to catalyse the production of ATP from ADP.

5.3. Secondary active transport

In some cases, it is advantageous for the cell to use a pre-existing gradient of a chemical species in order to move another species against its chemical gradient. One common instance is a transporter employing the Na^+ gradient already established by the Na^+/K^+ ATP-ase and moving Na^+ ions into the cell in order to facilitate the movement of another ion against its gradient. This is called **secondary active transport**, as the gradient used as an energy source has already been established in a separate process that expended ATP. Thus, in

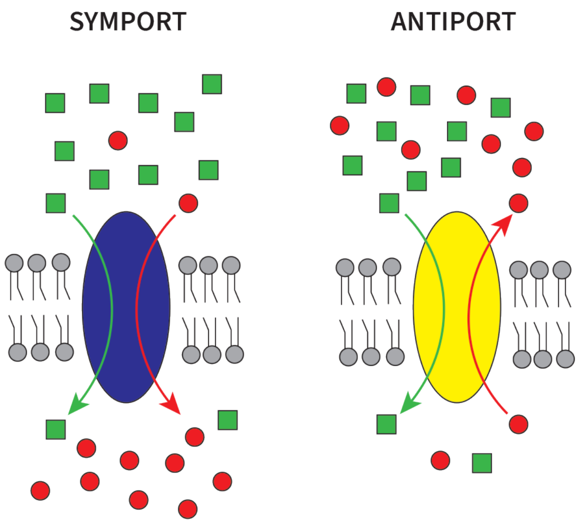


Figure 5.16. Secondary active transporters. Left – a symporter; right – an antiporter. One chemical species (squares) is transported according to its gradient in both cases; the other (circles) is transported against its gradient.

secondary active transport, two chemical species are moved at the same time: one in the direction of its gradient (from high to low potential) and the other against its gradient.

Two main types of secondary active transport exist, depending on whether the two species transported both move from the same side of the membrane to the other (**symport**) or one moves inside the cell while the other one exits the cell (**antiport**). These are shown in Figure 5.16.

We will mention only one example of transporter from each category. The sodium/glucose cotransporters (SGLT1 and SGLT2) use the Na^+ gradient established by the Na^+/K^+ ATP-ase in order to transport glucose against its concentration gradient. These are present in the small intestine and are used to ensure that all glucose is retained by the organism. As both species are transported from the intestinal lumen towards the cytoplasm of the enterocyte, SGLTs are symporters. A property of SGLT1 is its high cooperativity – binding of Na^+ causes affinity for glucose to increase more than 10 times.

The Na^+/Ca^{2+} antiporter (NCX) removes Ca^{2+} ions from cells by using the pre-existing Na^+ gradient, bringing Na^+ inside the cell. This ensures that the intracellular Ca^{2+} concentration is low, which is essential in many processes, for example in the contractile action of the heart muscle.

6. RECEPTORS

Membrane receptors are specialized membrane proteins that are involved in cell signaling processes. Receptors are able to bind to extracellular signaling molecules called first messengers

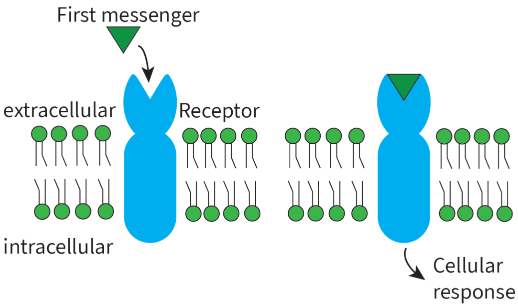


Figure 5.17. Basic mechanism of a receptor. An extracellular first messenger (triangle) binds to the receptor, which triggers a response from the cell.

(hormones, growth factors, neurotransmitters, etc.), and then produce a chemical response that induces changes in the activity of the cell. This chemical response can be a conformational or chemical change of the receptor itself or the production of another molecule called a second messenger. The basic function of a receptor is shown in Figure 5.17.

Three main categories of membrane receptors exist:

- ion channel linked receptors;
- G-protein coupled receptors;
- enzyme-linked receptors.

Examples of the first two are shown in other chapters of this book: the nicotinic acetylcholine receptor is an example of an ion channel linked receptor while rhodopsin is a G-protein coupled receptor. As an example of an enzyme-linked receptor we can mention the Epidermal Growth Factor Receptor (EGFR) that binds growth factors which can then trigger the autophosphorylation of residues on the receptor itself, resulting in downstream signaling by proteins that can bind to the receptor in its phosphorylated form.

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell 4th Edition: International Student Edition*. New York: Garland Science.

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Bazzone, A., Zerlotti, R., Barthmes, M., & Fertig, N. (2023). Functional characterization of SGLT1 using SSM-based electrophysiology: Kinetics of sugar binding and translocation. *Frontiers in Physiology, 14*. doi:10.3389/fphys.2023.1058583

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2012). *Biochemistry. Seventh Edition*. New York: Freeman and Company.

Bischofberger, M., Iacovache, I., & van der Goot, F. G. (2012). Pathogenic pore-forming proteins: function and host response. *Cell Host Microbe, 12*(3), 266-275. doi:10.1016/j.chom.2012.08.005

Boron, W. F., & Boulpaep, E. L. (2017). *Medical Physiology* (3 ed.). Philadelphia: Elsevier.

Carruthers, A., DeZutter, J., Ganguly, A., & Devaskar, S. U. (2009). Will the original glucose transporter isoform please stand up! *Am J Physiol Endocrinol Metab*, 297(4), E836-848. doi:10.1152/ajpendo.00496.2009

Dillon, P. F. (2012). *Biophysics: A Physiological Approach*. Cambridge: Cambridge University Press.

Glaser, R. (2012). *Biophysics: An Introduction*. Heidelberg: Springer.

Guyton, A. C., & Hall, J. E. (2005). *Textbook of Medical Physiology. Eleventh Edition*. Philadelphia: Elsevier.

Nelson, D. L., Cox, M. M., & Hoskins, A. A. (2021). *Lehninger Principles of Biochemistry Eighth Edition*. New York: MacMillan Learning.

Tang, M., & Monani, U. R. (2021). Glut1 deficiency syndrome: New and emerging insights into a prototypical brain energy failure disorder. *Neurosci Insights, 16*, 26331055211011507. doi:10.1177/26331055211011507

CHAPTER 6

BIOELECTRICITY

Prerequisite knowledge

- Electric charge. Coulomb's law
- Properties of membranes
- Function of ion channels

1. BASICS OF BIOELECTRICAL PHENOMENA

All living cells rely on bioelectrical phenomena in order to function and survive. These occur due to the presence of electrically charged chemical species inside and outside the cell. The distribution and movement of these chemical species is engineered by our cells in order to produce a state of electric charge imbalance across the cellular membrane called the membrane potential. Cells then rely on the existence of this membrane potential in order to properly function, respond to external stimuli and communicate with other cells.

As mentioned in a previous chapter, the membrane potential appears due to the asymmetrical distribution of the concentrations of certain ions such as Na^+ or K^+ across the membrane. In order to fully understand how the membrane potential is generated, we will employ a simple study system (Figure 6.1) that we will progressively alter to become more and more complex.

The model system can be described as follows. We take a container with two compartments, (1) and (2), that are separated by a biological membrane. If it is easier for you, think of this system from the beginning as analogous to the cell – compartment 1 is the cytoplasm and compartment 2 the extracellular medium, although it might not look like this initially. Each compartment can contain different concentrations of certain chemical species that are either permeable or impermeable through the membrane depending on the conditions we set. The solvent is water, as in a living organism. Finally, in each compartment a metal wire is inserted to serve as an electrode. These two electrodes are connected to a measurement instrument (voltmeter) capable of measuring the potential difference between the two compartments (the membrane potential).

1.1. Case 1. The model system using an impermeable membrane

If the membrane is impermeable (Figure 6.1), the system is easily described. We'll start with different concentrations of KCl in each compartment, with $c_1 > c_2$.

The compartments are initially electrically neutral, as $[K^+]_1 = [Cl^-]_1$ and $[K^+]_2 = [Cl^-]_2$. No ions can move across the membrane, therefore each compartment remains electrically neutral at all points in time.

There will be, thus, no imbalance of charges across the membrane, and the voltmeter will show a membrane potential of zero.

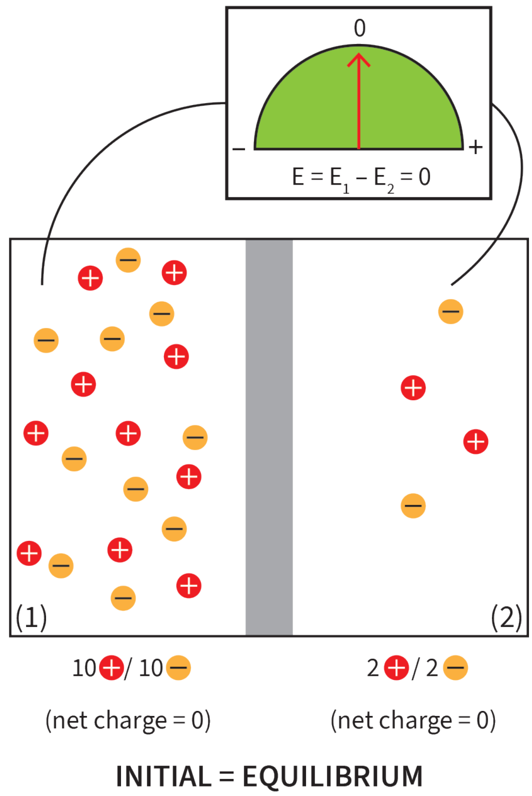


Figure 6.1. Case 1. The model system. In this initial example, compartment 1 contains a high concentration of K^+ (red circles) and Cl^- (orange circles), but both ions are impermeable through the membrane. In each compartment, an electrode is inserted and the membrane potential E can be measured by the voltmeter.

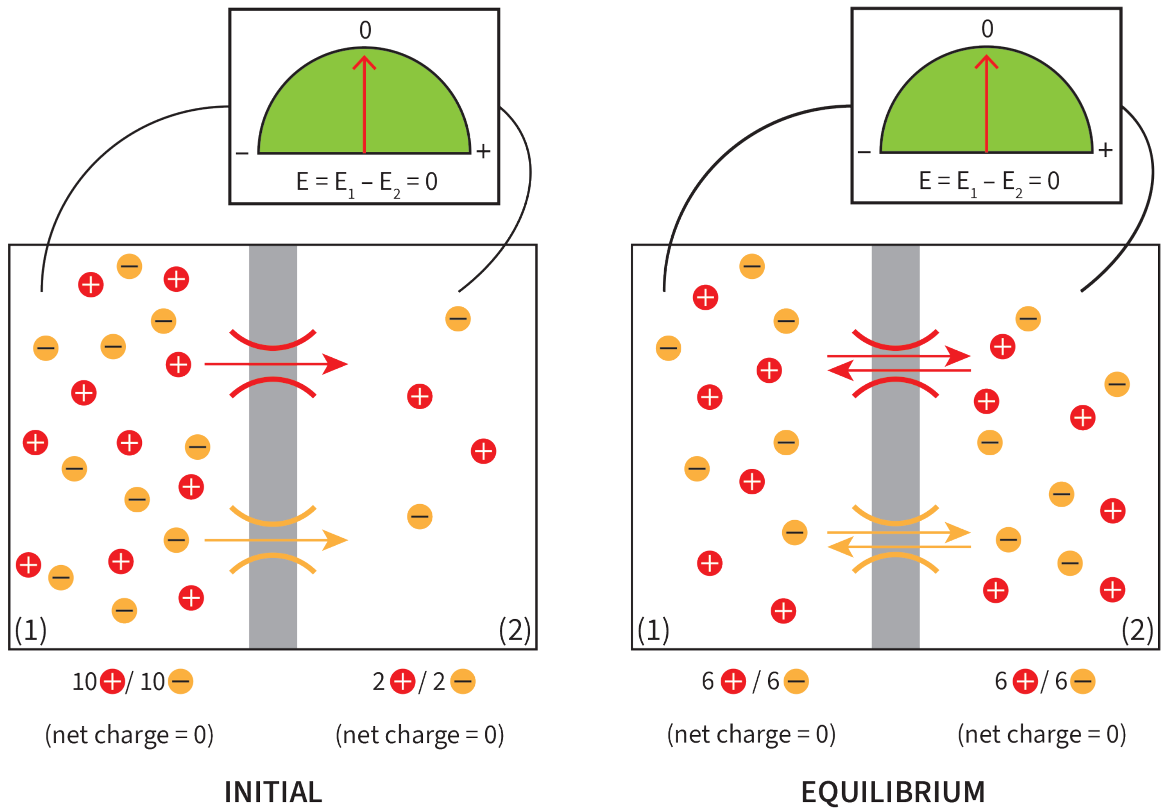


Figure 6.2. Case 2. The model system as in Figure 6.1, only that the membrane has been made equally permeable to both K^+ and Cl^-. Equilibrium is reached when concentrations of K^+ and Cl^- equalize across the membrane.

1.2. Case 2. The model system using a membrane equally permeable to both ions

To build the system in case 2, we start with the situation in case 1, but we make the membrane permeable to both ions (Figure 6.2). As we studied diffusion in the previous chapters, we should have an intuitive understanding of what will be happening. Both K^+ and Cl^- will diffuse through the membrane according to (we can also say "down") their concentration gradient. If their permeabilities are exactly equal, the same number of K^+ ions and Cl^- ions will diffuse over time, keeping each compartment neutral.

Transport will stop when equilibrium is reached, thus when the concentration of both ions in compartment 1 is equal to that in compartment 2. Using the same starting number of ions as in the previous example (10 ions of each in compartment 1 and 2 in compartment 2), we can easily see that equilibrium is reached when we have 6 ions of K^+ in both compartments 1 and 2, and 6 ions of Cl^- in compartments 1 and 2. As the net charge in both compartments is, again, 0, the membrane potential measured by the voltmeter will also be 0.

1.3. Case 3. The model system using a membrane permeable to only one ion

This example (Figure 6.3) is key to understanding the first part of this chapter. Let's pay special attention to what happens if we make the membrane permeable only to K^+. As before, we know K^+ ions will diffuse from left to right. However, Cl^- ions cannot pass through the membrane. Despite what our intuition tells us, **equilibrium is not reached when K^+ concentrations equalize**. Instead, equilibrium is reached much, much earlier, due to the appearance of an electrical gradient that will act in the opposite sense to the concentration gradient and will eventually stop net transport of ions from left to right.

Let's slowly analyze what is happening. In the previous example, we did not really care about charge, as always, the movement of one positive charge (one K^+ ion) was accompanied by the movement of one negative charge (one Cl^- ion).

When we make only K^+ move, we quickly start to create an imbalance of charges. Let's consider ion movement from compartment 1 to compartment 2, using some larger numbers than what we used before (Figure 6.4). Thus, let's say we started with 500 ions of K^+ and Cl^- in the left compartment, and 20 ions of K^+ and Cl^- in the right compartment.

As soon as only one K^+ moves we have a net charge of 499–500 = –1 in the left compartment and 21–20 = +1 in the right compartment. This has now created an electrical gradient between the two compartments (remember that same charges repel, opposite charges attract!). There are, now,

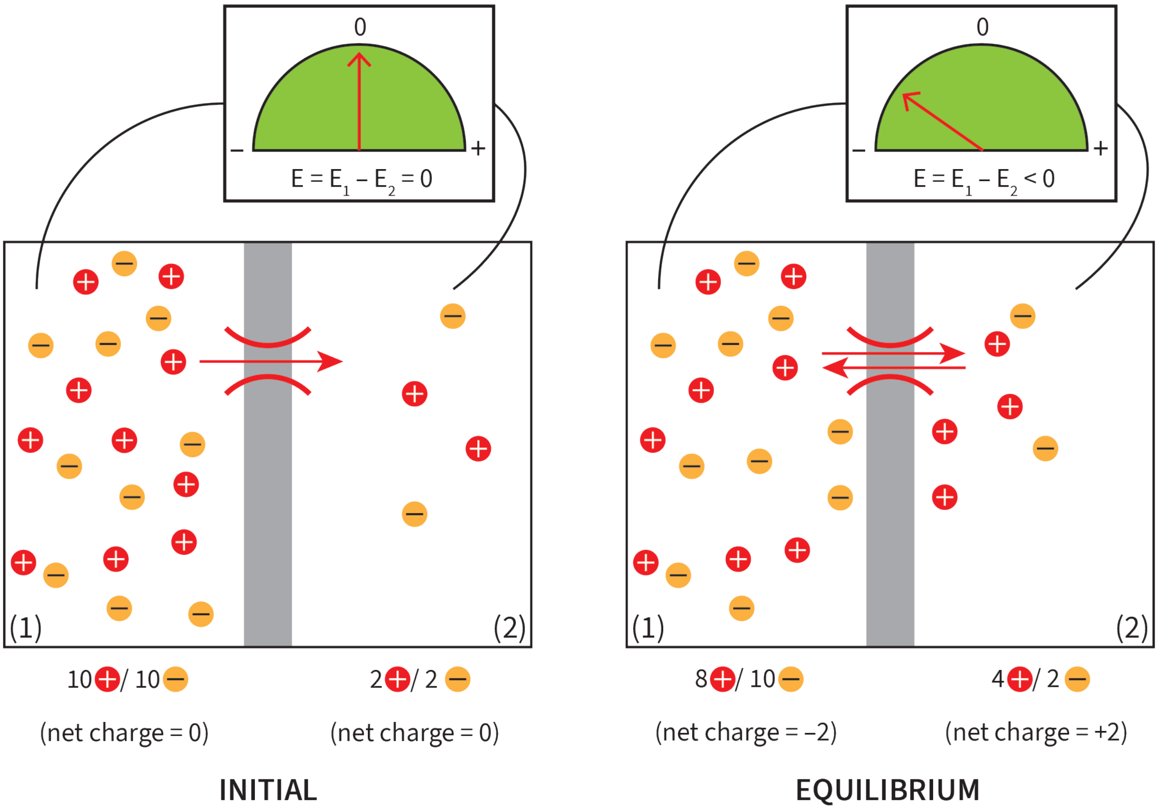


Figure 6.3. Case 3. The model system as in Figure 6.1, only that the membrane has been made only permeable to K^+ and not to Cl^-. At equilibrium, the concentrations of K^+ across the membrane are not equal.

two types of gradients in the system. The concentration gradient will continue to want to move K^+ from left to right, but the electrical gradient will want to move K^+ from right to left! We can calculate the **net transport** from compartment 1 to compartment 2 as the difference between the flow of diffusion from 1 to 2 (caused by the concentration gradient) and the flow of diffusion from 2 to 1 (caused by the electrical gradient).

Using our arbitrary numbers, let's say that the electrical gradient after a single K^+ ion has moved is not yet as large as the concentration gradient, as shown in the left panel of Figure 6.4. However, the more ions of K^+ move from left to right, the more the electrical gradient repelling them from the right compartment increases, until at some point it becomes as large as the concentration gradient. Let's say this happens after 5 ions of K^+ have moved from left to right. From now on, even if the concentration gradient drives more ions from left to right, quickly the electrical gradient will drive the same number back from right to left. We say that we have reached a state of equilibrium.

You are free to be surprised that equilibrium is not reached when concentrations are equal! Sometimes our intuition can be wrong... Instead, even a small buildup of electrical charge will soon be enough to counteract the concentration gradient. Note that the concentration gradient has not even changed that much. We started with 25 times more K^+ in the left compared to the right, and by the time equilibrium is reached there are still 20 times more K^+ ions in compartment 1 than in 2.

The numbers we gave in Figure 6.4 are small and easy to work with – they do not reflect the actual numbers in the cell, obviously. In a real

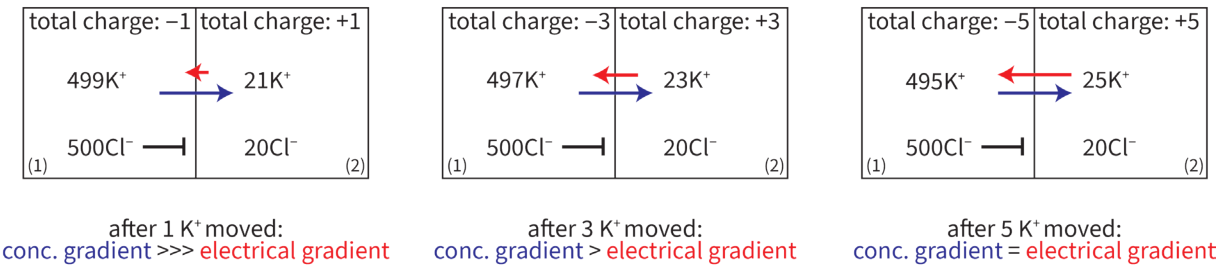


Figure 6.4. The model system as in Figure 6.3, evolving over time. We have chosen an example in which 500 K^+ ions and 500 Cl^- ions are initially in compartment 1, while 20 K^+ ions and 20 Cl^- ions are initially in compartment 2. The system reaches equilibrium (right panel) after enough K^+ ions have diffused from left to right.

cell, the equilibrium would be reached with only a negligible change in the concentration gradient, which is why we kept the blue arrow in Figure 6.4 the same size over time. In fact, in a real cell, less than 0.00001% of the total K^+ ions need to move over the membrane in order to establish the membrane potential.

Let's analyze the situation mathematically as well, using what we learned about the electrochemical potential in the previous chapter. Net transport will stop, and thus equilibrium will be reached when the electrochemical potential of K^+ in compartment 1 will be equal to the electrochemical potential of K^+ in compartment 2, or otherwise said:

$$\Delta W = W_1 - W_2 = 0 \tag{6.1}$$

Then, according to equation (5.4) from the previous chapter, and noting with c_1 and c_2 the concentrations of K^+ in each compartment:

$$RT\ ln\ \frac{c_1}{c_2} + zF(E_1 - E_2) = 0 \tag{6.2}$$

If we note $E_1 - E_2 = E_{eq}$, which is the membrane potential established at equilibrium (what our voltmeter will show) and $c_{ex} = c_2$, $c_{in} = c_1$ (remember we said in the beginning that compartment 1 is similar to the inside of the cell and 2 to the outside):

$$E_{eq} = \frac{RT}{zF}\ ln\ \frac{c_{ex}}{c_{in}} \tag{6.3}$$

where z is the electrical charge of the ion, F is Faraday's constant, R is the constant of ideal gases and T is the temperature (in K).

Equation (6.3) is called the **Nernst equation**, as mentioned in the previous lecture. It allows us to calculate the equilibrium potential of an ion, which is the membrane potential at which the net movement of a particular ion stops. The **equilibrium potential** also has two other names that can be used interchangeably: **Nernst potential** or **reversal potential**.

Using the Nernst equation as written in equation (6.3) and replacing the values of the constants in the International System we will get the result in V (volts).

A simplified form of the Nernst equation replaces the natural logarithm with a decimal logarithm and uses standard conditions of temperature (T = 298 K) to quickly provide a value of the membrane potential in mV (millivolts) at 298K:

$$E_{eq}(mV) = \frac{59}{z}\ lg\ \frac{c_{ex}}{c_{in}} \tag{6.4}$$

Note that it only makes sense to calculate a value of the Nernst potential for ions that are permeable through the membrane, so in our previous example, only for K^+ but not for Cl^-. **If only one ion**

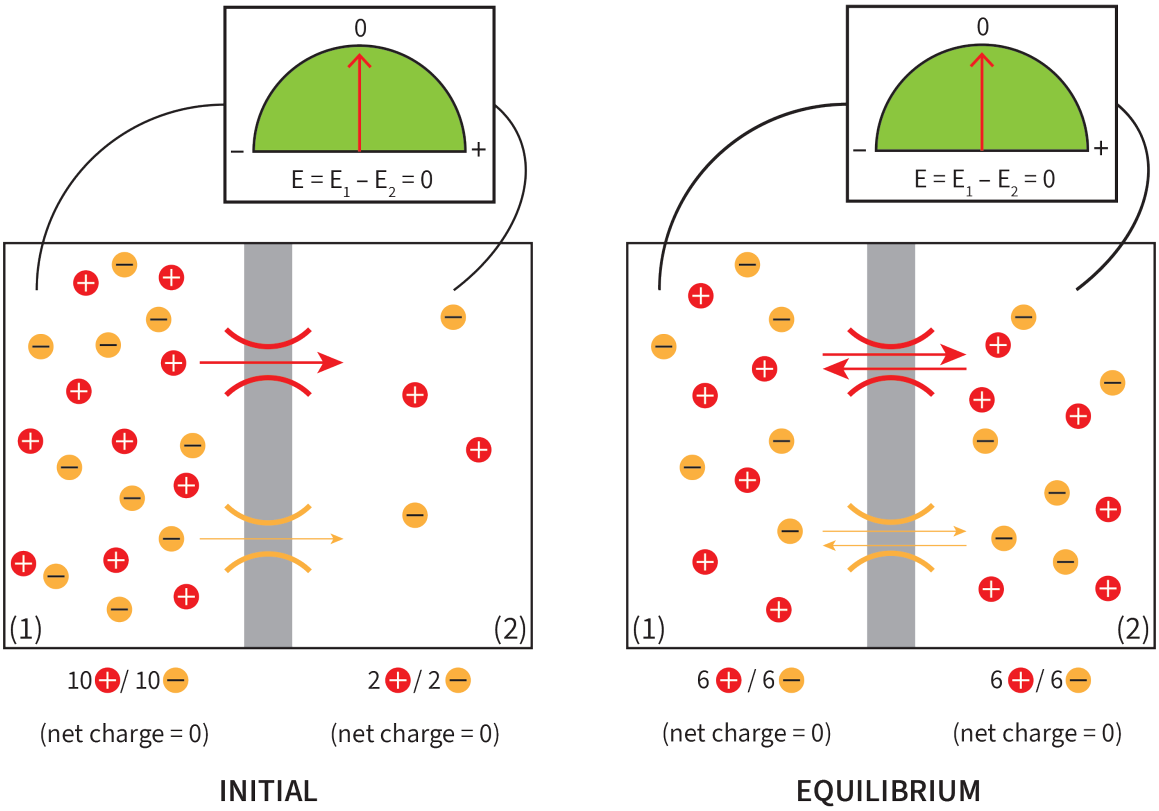


Figure 6.5. Case 4. The model system with both Cl^- and K^+ permeable, but with a lower permeability of Cl^-.

is permeable through the membrane, then the value of the membrane potential will be equal to the Nernst potential of that ion! It might be a little counterintuitive that the concentration of Cl^- does not directly affect the value of the membrane potential, but we will see a similar example later.

1.4. Case 4. The model system using a membrane unequally permeable to both ions

What if Cl^- is now a little bit permeable (though still less than K^+), as in Figure 6.5? Note that in a typical cell the permeability of Cl^- is about half that of K^+.

This time, we should relatively easily know what will happen both before and at equilibrium. Before equilibrium we should be in a situation relatively similar to what is happening in Case 3, whereas equilibrium will be reached at some point when concentrations equalize, as in Case 2. There is no membrane potential at equilibrium, and before equilibrium we can calculate the membrane potential with the following equation, called the Planck-Henderson equation:

$$\Delta E = E_1 - E_2 = \frac{P_K - P_{Cl}}{P_{Cl} + P_K} \cdot \frac{RT}{zF} \ln \frac{c_2}{c_1} \qquad (6.5)$$

We say that, before equilibrium, a diffusion potential is established that slows the faster ion (K^+) and speeds up the slower ion (Cl^-). We will not use the Planck-Henderson equation further, but notice that, if $P_{Cl} = 0$, as in case 3, equation (6.5) turns into the Nernst equation.

1.5. Case 5. The model system adding an impermeant large anion. The Donnan equilibrium

We are moving closer to making our model more and more like a real cell. In the cell, there are proteins that are negatively charged at physiological pH. Let's, then, add such an impermeant large anion to compartment 1 which will replace Cl^- (Figure 6.6). We will start this time with equal concentrations of K^+ across the membrane, and with overall electroneutral compartments (the number of positive charges equals that of negative charges in both compartments 1 and 2). Finally, we will make the membrane semi-permeable (permeable to the solvent, water), which is a general property of biological membranes.

In the system of Figure 6.6, it is obvious that Cl^- will start to diffuse from compartment 2 to compartment 1, building a negative charge across the left side of the membrane. This will, then, cause the diffusion of K^+ from compartment 2 towards compartment 1 in order to compensate this charge. Net transport of both ions will stop once the membrane potential is equal to the Nernst potential of both permeant ions, K^+ and Cl^-. Remember that **the impermeant anions do not contribute to the resting membrane potential.** We have, then:

$$\Delta E = \frac{RT}{F} \ln \frac{[K^+]_2}{[K^+]_1} = -\frac{RT}{F} \ln \frac{[Cl^-]_2}{[Cl^-]_1} \qquad (6.6)$$

where the minus sign in the second term comes from $z = -1$ for Cl^-.

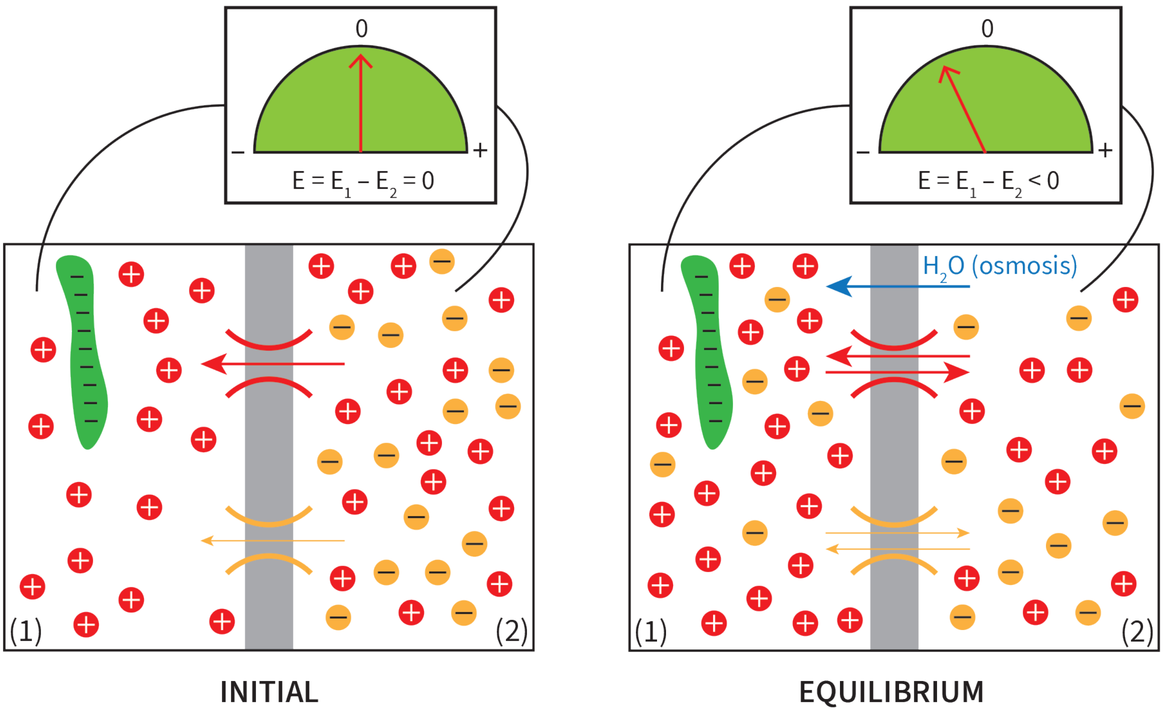


Figure 6.6. Case 5. The model system with impermeant large anions (green shape) added to compartment 1. Initially, the charge in both compartments is 0 and $[K^+]_1 = [K^+]_2 = [Cl^-]_2$.

We can say that, in order to equalize the electrochemical potentials of the K^+ and Cl^- ions, a particular distribution of their concentrations is established, called a **Donnan equilibrium**. The Donnan equilibrium can be described by the following equation:

$$\frac{[K^+]_2}{[K^+]_1} = \frac{[Cl^-]_1}{[Cl^-]_2} \tag{6.7}$$

The Donnan equilibrium is not, however, a true state of thermodynamic equilibrium, as the presence of the protein anions inside compartment 1 exerts a very high colloid osmotic pressure. Thus, water will enter compartment 1 through osmosis. In a real cell, this would cause the inside of the cell to swell until the membrane bursts. To get to a model of a real cell we have to add one more ion to compartment 2 in order to compensate this osmotic flow.

1.6. Case 6. Making our model as close to the cell as we can by adding Na^+

The final step is to add Na^+ to the equilibrium state of case 5 in order to prevent the osmotic flow of water. To make Na^+ stay in (2), we will make it impermeant through the membrane. In a real cell, Na^+ still has some permeability through the plasma membrane, but it is much lower (about 25 times) than that of K^+. The final distribution of the ions in the cell looks similar to Figure 6.7 at equilibrium.

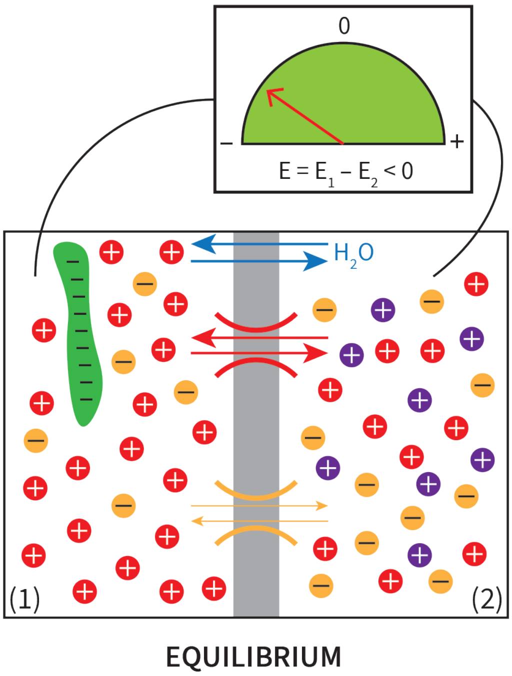


Figure 6.7. Case 6. The model system similar to a real cell. Na^+ (purple circles) was added to the right compartment in order to balance the osmotic pressures of both compartments. Remember that compartment 1 was analogous to the intracellular side of the membrane and compartment 2 to the extracellular side. At equilibrium, $[K^+]_1 > [K^+]_2$, $[Na^+]_1 < [Na^+]_2$ and $[Cl^-]_1 < [Cl^-]_2$. Protein anions (green shape) are only found in (1).

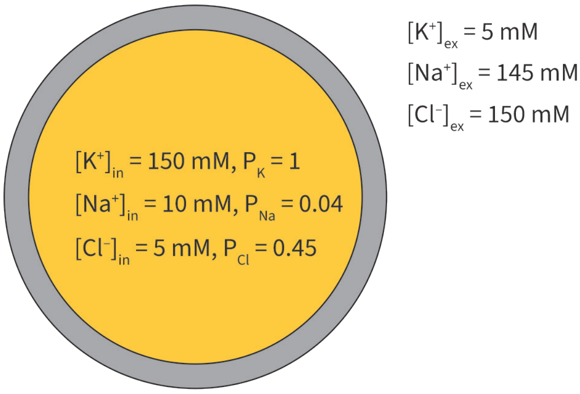


Figure 6.8. Distribution of ions across the membrane of a typical cell. Listed are the concentrations and permeabilities through the membrane. Concentrations and permeabilities are given as examples and can differ depending on the cell type.

2. THE RESTING MEMBRANE POTENTIAL

2.1. Resting potential of a real cellular membrane

Let us recap what we learned before and put it into the context of a real cell. An asymmetrical distribution of ions across the membrane is established by the cell. This is maintained actively, in the case of Na^+ and K^+, through the action of the Na^+ and K^+ ATP-ase, and passively, in the case of Cl^- (through Cl^- channels). Figure 6.8 shows an example of concentrations for these chemical species inside and outside a typical cell. Note that this can, and does vary between different cell types.

What we call the **resting membrane potential** is established: the inner side of the membrane is more negative than the outer side **and the value of the resting membrane potential is usually between –50 mV and –100 mV** depending on the type of cell. The resting membrane potential is a characteristic of all living cells.

We also say that a cellular membrane that has a negative-inside membrane potential is *polarized* (there is a separation of charges between the two sides of the membrane).

2.2. The Goldman-Hodgkin-Katz equation

Let's calculate the Nernst potentials of ions in Figure 6.8 using the Nernst equation (equation 6.4). We obtain: $E_K = -87$ mV, $E_{Cl} = -87$ mV and $E_{Na} = +69$ mV. Having said before that the usual value for the membrane potential in a typical cell is between –50 mV and –100 mV, we can conclude

that the value of the membrane potential is much closer to the Nernst potential of K^+, but not to that of Na^+. Obviously, this is also close to the Nernst potential of Cl^-, but, since Cl^- is generally not actively transported, we don't consider it to be as important as K^+.

If we want to be very precise we can use an equation called the Goldman-Hodgkin-Katz equation (or just Goldman equation in short) is order to calculate the membrane potential if we know the concentrations and permeabilities of the most important ions in the cell. Generally, this can be written as:

$$\Delta E = \frac{RT}{F} \ln \frac{\sum P_C [C^+]_e + \sum P_A [A^-]_i}{\sum P_C [C^+]_i + \sum P_A [A^-]_e} \qquad (6.8)$$

where C^+ are the cations in the cell and A^- are the anions, while P_C and P_A respectively are the permeabilities of these species through the membrane. *e* and *i* refer to the extracellular and intracellular concentrations, respectively.

Notice that, if a particular ion is not permeable through the membrane, P = 0 for that ion, thus its concentration does not directly influence the value of the membrane potential.

To make the Goldman equation simpler, we can rewrite it only for Na^+, K^+ and Cl^- as:

$$\Delta E\ (V) = \frac{RT}{F} \ln \frac{P_K[K^+]_e + P_{Na}[Na^+]_e + P_{Cl}[Cl^-]_i}{P_K[K^+]_i + P_{Na}[Na^+]_i + P_{Cl}[Cl^-]_e} \qquad (6.9)$$

or, if T = 298K and converting to a decimal logarithm, we get the result in mV:

$$\Delta E\ (mV) = 59 \lg \frac{P_K[K^+]_e + P_{Na}[Na^+]_e + P_{Cl}[Cl^-]_i}{P_K[K^+]_i + P_{Na}[Na^+]_i + P_{Cl}[Cl^-]_e} \qquad (6.10)$$

If we input into equation (6.10) the values we gave in Figure 6.8, we obtain E = –69 mV which is, as we expected, close to the Nernst potential of K^+ that we calculated as –87 mV, but shifted to slightly more positive values by the influence exerted by the movement of Na^+ across the membrane.

We can also experimentally measure the value of the membrane potential for a particular cell. This is done by inserting a glass pipette with a very fine tip inside the cell. Inside the pipette a metal electrode is inserted (usually a silver wire coated with silver chloride) and the pipette is filled with KCl solution. Outside the cell, a second electrode is placed and the two are connected to a measurement device (voltmeter or oscilloscope).

3. ACTION POTENTIALS

3.1. Excitable cells. General concepts

For most cells, the value of the membrane potential is kept at a relatively constant value throughout their life. However, a few types of cells: **neurons, muscle cells and some endocrine cells** (such as the β-cells that secrete insulin in the pancreas) have the capability of performing rapid and potentially drastic changes in their membrane potential when subjected to an external stimulus; they are called **excitable cells**.

Let us first introduce some terms that we will use further. As we saw before, we call a membrane that has a normal, negative-inside resting potential, to be **polarized**. Note that, as previously mentioned, all cells rely on the existence of a resting membrane potential, no matter whether they are excitable or non-excitable. As a general rule, non-excitable cells have slightly less negative membrane potentials than excitable cells, but their membrane is still polarized (for instance, the resting potential of red blood cells is ~ –10 mV).

If the membrane potential increases from the resting value to less negative or even positive values, we say that the membrane was **depolarized**. If a membrane was first depolarized, and the membrane potential then returns to the resting value, we say that it was **repolarized**. Finally, if the membrane potential drops to more negative values than the resting potential, we say that the membrane was **hyperpolarized**.

The change in the membrane potential of an excitable cell occurs as response to a stimulus – this can be any kind of change in the external conditions of a cell: modification of the concentration of a certain ion or ligand, arrival of an action potential from another cell linked by a synapse, absorption of photons, etc. Not all stimuli will trigger a response from an excitable cell, of course. For example, only photoreceptor cells are specialized to respond to an arriving photon and trigger a biochemical cascade. Additionally, even if a cell is capable of responding to a certain type of stimulus, the intensity of the stimulus also has to be considered. If the stimulus is of low intensity, it will only trigger a limited response called a **local potential** (also known as a graded potential). Depending on the type of stimulus, local potentials can be either depolarizing or hyperpolarizing. Their intensity depends on the strength of the stimulus (hence the name graded potentials).

However, if the stimulus is depolarizing and is strong enough to trigger a sufficiently high depolarization in the excitable cell that goes above a certain limit (called the firing threshold), the response of the excitable cell will be maximal (called an **all-or-none action potential**).

3.2. Local potentials. All-or-none action potentials

To sum up the last paragraphs: a *local potential*

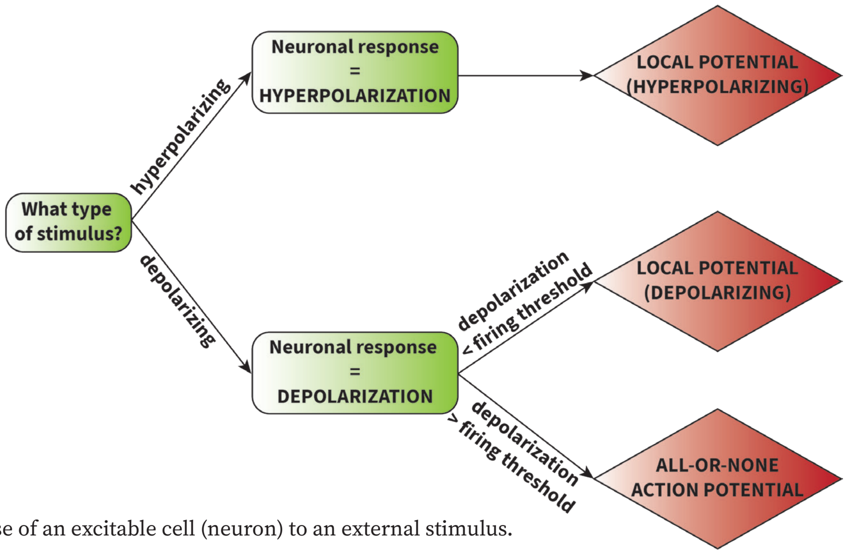


Figure 6.9. Response of an excitable cell (neuron) to an external stimulus.

(*LP*) is a limited response of an excitable cell to a low-intensity stimulus while an *all-or-none action potential* (*AP*) is a maximal response to a stimulus strong enough to induce depolarization to the firing threshold (Figure 6.9). To use an analogy, a local potential is like slowly turning up the volume of your favorite music player, while triggering an all-or-none action potential is like flipping a light switch: either the light is off, or on, with nothing in between. For the sake of simplicity, note that, from now on, we will usually shorten "all-or-none action potential" to just "action potential".

As their name suggests, local potentials stay local; they are triggered at a specific point on the cellular membrane and their propagation to neighboring parts of the membrane occurs with loss (*decremental* or *electrotonic propagation*). Therefore, they attenuate quickly and cannot affect the entire membrane of a cell. For example, in neurons, local potentials are triggered in the dendrite membrane. Unlike local potentials, action potentials propagate without loss and affect the entire membrane of the excitable cell. The main differences between local potentials and action potentials are summarized in Table 6.1.

3.3. Phases of local and action potentials

Let us analyze the general shape of local and action potentials by following the evolution of the membrane potential over time (Figure 6.10). In this example, we will only consider the response to a depolarizing stimulus.

The excitable cell detects a depolarizing stimulus and, depending on its intensity, either triggers a local or an action potential. In both cases, an initial, ascending phase consists of the depolarization of the membrane. The final value of the membrane potential after depolarization is modulated by the intensity of the stimulus in the case of an LP, and is constant (maximal) for an AP. The ascending phase is quickly followed by a descending phase where the membrane potential returns towards the normal, resting value. In the case of a neuronal AP, there is a short period of hyperpolarization, where the membrane potential drops below the resting value. Overall, for most excitable cells, an AP lasts for a few (1 – 5) milliseconds, but there are exceptions, such as the cardiac muscle cells, where the duration of the AP is much longer (hundreds of ms). During the triggering of an AP, the excitable cell enters a refractory period, where it is less sensitive to additional external stimuli. Thus, during the absolute refractory period, any additional stimuli sensed by the cell will no longer trigger a response, as the cell is already undergoing an AP. During the relative refractory period, the cell can still respond to external stimuli, but these have to be more intense than normal to trigger another AP.

Table 6.1. Differences between local and action potentials.

	LOCAL POTENTIAL	ACTION POTENTIAL
Trigger	Hyperpolarizing stimulus or weak depolarizing stimulus	Strong depolarizing stimulus
Type of response	Hyperpolarizing or depolarizing	Depolarizing
Amplitude	Depends on the intensity of the stimulus	Always maximal
Propagation	Decremental (with loss)	Non-decremental (lossless)

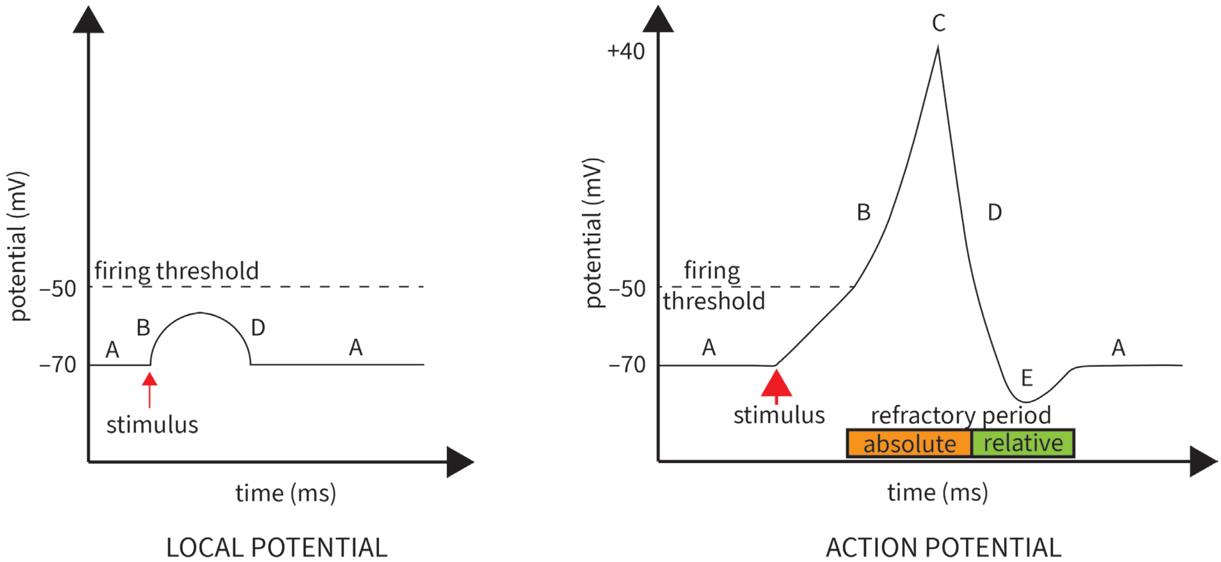


Figure 6.10. Phases of a depolarizing local potential (left) and an action potential (right). Initially the membrane is at rest (A) and a stimulus triggers an ascending phase of depolarization (B). If the stimulus is intense enough and the depolarization goes above the firing threshold, an action potential is triggered, reaching a peak potential (C). This is followed by the descending phase where the membrane is repolarized (D). In the case of an AP, there is also a phase of hyperpolarization (E) before the membrane potential is restored to the resting value (A).

3.4. Molecular events

What triggers these rapid changes in the membrane potential? The answer is the transport of ions over the membrane through ion channels. Ion channels are particularly adapted to facilitate the appearance of LPs and APs because they are:

- fast (a single channel can facilitate the movement of more than one million ions per second);
- tightly regulated by the existence of gates that will only open when a specific stimulus activates their sensor region.

In most excitable cell, two main types of ion channels are involved in triggering LPs and APs: Na^+ channels and K^+ channels. We said that, under resting conditions, the permeability of K^+ through the cellular membrane is much higher than that of Na^+. This occurs because of the existence in the membrane of the so-called K^+ "leak channels", that remain open under resting conditions. The K^+ leak channels are not the only type of K^+ channel that exists in the cell, as there are also additional, voltage-gated K^+ channels, that are activated by an increase (depolarization) of the membrane potential. Schematically, these can be represented as in Figure 6.11.

Na^+ channels also exist in excitable cells. Two main types of Na^+ channels are involved in triggering LPs and APs: ligand-gated Na^+ channels and voltage-gated Na^+ channels. Unlike K^+ channels or ligand-gated Na^+ channels, which have a single gate, voltage-gated Na^+ channels have two gates: an activation gate and an inactivation gate (Figure 6.12). When either one or both of these gates are closed, no ions can pass through the channel. When the activation gate is closed, we say that the channel is closed; when the inactivation gate is closed we say that the channel is inactivated.

Let's follow the molecular events occurring during the triggering of an AP from start to finish (Figure 6.13). Initially, an external stimulus triggers an influx of cations which depolarizes the

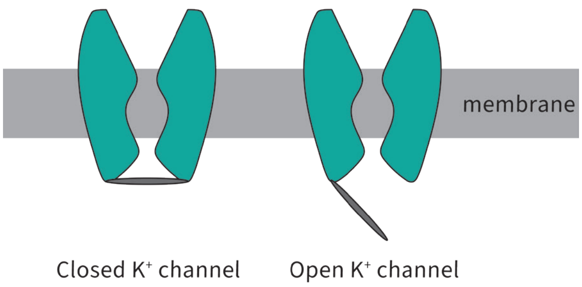


Figure 6.11. Voltage-gated K^+ channels. Under resting conditions, the gate of these channels is closed (left), but opens (right) when the membrane potential increases above a certain limit.

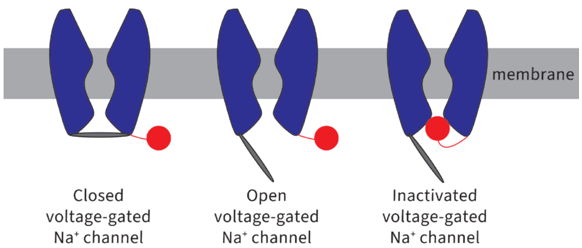


Figure 6.12. Voltage-gated Na^+ channels. Under resting conditions, the activation gate of these channels is closed (left), but opens (middle) when the membrane potential increases above a certain limit. After a short time has passed, the channel is inactivated (right) by closing the inactivation gate.

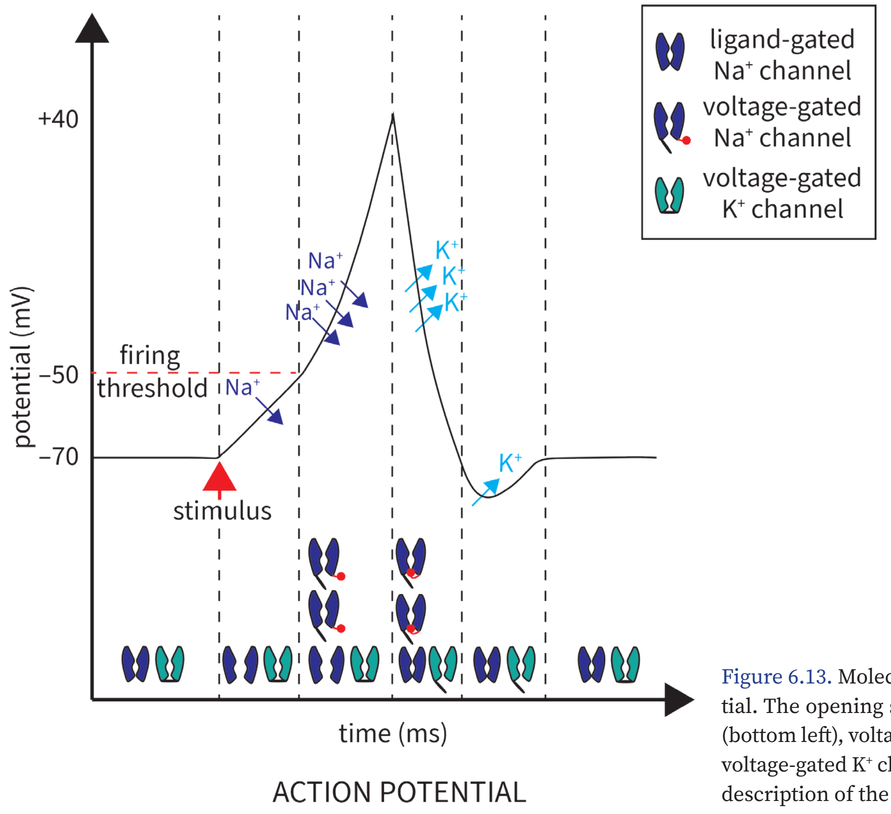


Figure 6.13. Molecular events during an action potential. The opening states of ligand-gated Na^+ channels (bottom left), voltage-gated Na^+ channels (top left) and voltage-gated K^+ channels (right) are shown. Detailed description of the events is provided in the text.

membrane. This can occur, for example, through a ligand-gated Na^+ channel, a Ca^{2+} channel, etc. We'll use a ligand-gated Na^+ channel for the example in Figure 6.13.

If just a small number of these cation channels open, only a local potential is triggered. If many of them open, and the membrane potential reaches the firing threshold, a cascade effect starts to occur. Voltage-gated Na^+ channels will open, making the membrane potential even more positive, and causing the opening of even more voltage-gated Na^+ channels. Fortunately, this does not go on too long, as a continuous state of high depolarization would have extremely detrimental effect to the cell. The voltage-gated Na^+ channels are relatively rapidly inactivated by the time the membrane potential reaches a peak value that is usually around +40 mV. Thus, voltage-gated Na^+ channels are only open for a very short period of time. The depolarization will also cause voltage-gated K^+ channels to open; K^+ will exit the cell driven by its high concentration gradient and the now inside-positive membrane potential. This quickly restores the membrane potential down to more negative values. Usually, these K^+ channels stay open a little longer, therefore causing a small "overshoot" of hyperpolarization. Once these close, though, the Na^+/K^+ ATP-ase that functions continuously throughout this process will restore the membrane potential to its normal resting value.

Note that these ion movements do not really change the bulk concentration of ions inside the cell, as they only occur for such a small amount of time (a few ms). Only the local concentrations very close to the membrane are changed, but this is enough to significantly shift the membrane potential value.

We can also analyze the events that happen from the point of view of the Nernst potentials. At rest, only K^+ leak channels are open and the permeability of K^+ is much higher than that of Na^+. Therefore, the membrane potential stays close to the Nernst potential of K^+ (~ −80 mV). As soon as enough Na^+ channels open, the permeability of Na^+ through the membrane increases significantly, and the membrane potential moves towards the Nernst potential of Na^+ (~ +60 mV). When the Na^+ channels are inactivated, the membrane potential drops again towards the Nernst potential of K^+.

4. PROPAGATION OF ACTION POTENTIALS IN THE CELLULAR MEMBRANE

If an AP or LP is generated in a particular point on the cellular membrane, this will cause a local modification in the distribution of charges across the membrane. This results in a migration of electric charges (electric currents) towards neighboring regions. In the case of an action potential, this will cause the entire membrane of

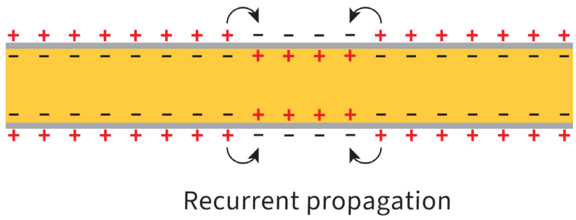


Figure 6.14. Recurrent propagation in a non-myelinated nerve fiber. A portion of the membrane (center) is depolarized when an AP is triggered. The neighboring regions will soon also be depolarized as the depolarization propagates through the fiber.

the cell to be depolarized. In the case of a local potential, this is quickly attenuated. There is one exception: if several depolarizing local potentials are generated in close proximity to each other, the sum of their effects might be enough to bring the membrane potential to the firing threshold and thus trigger an action potential.

Two main mechanisms are responsible for the propagation of APs. In most cells, this is performed by a process called recurrent propagation. This is a slow process, as it requires every neighboring region of the membrane to be depolarized to the firing threshold, which, in turn, depolarizes its neighbors, and so on. Propagation speed is limited at 0.5 – 2 m/s. This process happens, for instance, in non-myelinated nerve fibers (Figure 6.14).

In order to increase the propagation speed of action potentials, most nerve fibers in our body are covered by an insulating mixture of lipids and proteins called the myelin sheath. Myelin is synthesized by glial cells: by oligodendrocytes in the central nervous system and by Schwann cells in the peripheral nervous system. The myelin sheath does not cover the entire nerve fiber, but leaves some openings where the plasma membrane is exposed to the extracellular medium, called Nodes of Ranvier (Figure 6.15). The regions covered by myelin are called internodes. As myelin is an electrical insulator, the membrane is not exposed to the extracellular membrane at the internodes. The propagation of the AP can, thus, skip the internodes and will "jump" directly to the neighboring nodes of Ranvier. This is called saltatory propagation. In this manner, the action potential is generated only at the nodes, and the propagation of the AP is much faster: in myelinated nerve fibers, propagation is up to 100 times faster than in non-myelinated fibers.

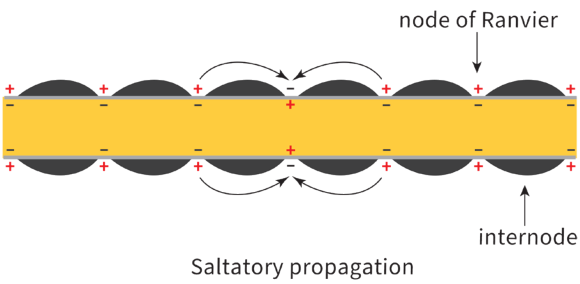


Figure 6.15. Saltatory propagation in a myelinated nerve fiber. A portion of the membrane (center) is depolarized when an AP is triggered. The neighboring nodes of Ranvier will be depolarized, while the membrane of the internodes is electrically insulated from the exterior.

As our body relies on the myelin sheath, diseases that affect it are tremendously debilitating and potentially lethal. The most common demyelinating disease is multiple sclerosis, a chronic condition in which the myelin sheath is damaged by the body's own immune system. At the moment, there is no cure for this disease.

5. SYNAPSES

We have seen how the action potential is triggered in a cell and how it can then propagate through the entire membrane of that cell. How can the action potential be transmitted to other cells? This is performed in our body by specialized structures called *synapses*.

Two main types of synapses exist in our bodies: **chemical synapses** (the vast majority) and **electrical synapses**. Let us describe each, in turn.

5.1. Chemical synapses

A chemical synapse facilitates the propagation of the AP from one excitable cell to another using as intermediary a chemical molecule called a *neurotransmitter*. Different types of neurons have their own, specialized, neurotransmitters. We will list here just a few: acetylcholine, glutamate, serotonin, dopamine, etc. The components of a chemical synapse are shown in Figure 6.16.

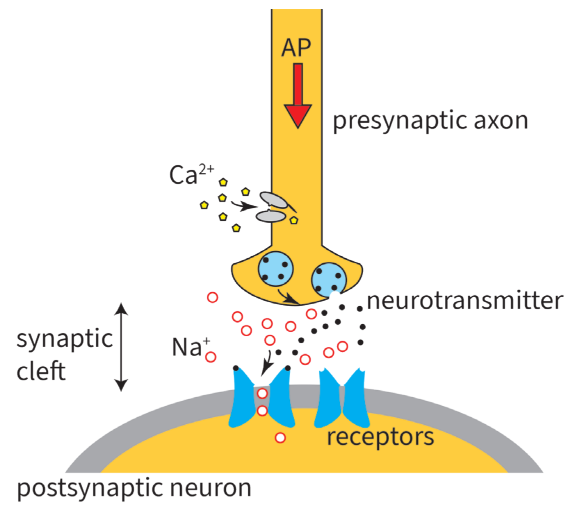


Figure 6.16. Example of an excitatory chemical synapse. Neurotransmitter molecules are shown by filled, black circles, Na^+ ions are empty circles and Ca^{2+} ions are yellow pentagons. A detailed description of the synaptic transmission can be found in the text.

Table 6.2. Differences between chemical and electrical synapses.

	Chemical synapses	Electrical synapses
Distribution in the body	Very common	Rare
Use neurotransmitters	Yes	No
Direction	Unidirectional (presynaptic to postsynaptic)	Bidirectional
Transmission speed	Slower	Fastest

A chemical synapse is formed between a presynaptic excitable cell and a postsynaptic excitable cell. Between the two there is a space of ~20 – 50 nm called the synaptic cleft. Transmission of the electrical signals (APs) in a chemical synapse is always unidirectional: the presynaptic cell contains vesicles filled with neurotransmitter molecules, while the membrane of the postsynaptic cell has receptors for that neurotransmitter. When an AP arrives at the synapse from the presynaptic cell, this causes voltage-gated Ca^{2+} channels in the membrane of the presynaptic cell to open. Following a chemical cascade, the vesicles filled with neurotransmitter fuse to the membrane, releasing neurotransmitter molecules in the synaptic cleft. These will diffuse across the synaptic cleft and bind to their receptors, modulating the activity of the postsynaptic cell. For example, if the synapse is excitatory, such as in Figure 6.16, activation of the receptors will cause excitation (depolarization) of the postsynaptic membrane which can be achieved, for instance, by opening of Na^+ channels in the postsynaptic membrane. If the LPs that result from this depolarization are large enough, this might trigger an AP in the postsynaptic cell. Note that the respective channel can be the receptor itself, as in Figure 6.16! By contrast, in an inhibitory synapse, arrival of an AP at the presynaptic cell will inhibit (hyperpolarize) the postsynaptic membrane.

An example of an excitatory synapse exactly like that in Figure 6.16 is the synapse between a motor neuron (presynaptic cell) and a muscle fiber (postsynaptic cell), called the neuromuscular junction. The neurotransmitter of this synapse is acetylcholine, and the receptor on the muscle fiber membrane is the nicotinic acetylcholine receptor, a ligand-gated ion channel (ionotropic receptor) that can transport Na^+ inside the cell when its gate is open by binding of acetylcholine.

In any case, after a short amount of time the neurotransmitter molecules will either be re-uptaken by the presynaptic cell, will diffuse out of the synaptic cleft or will be destroyed by enzymes in the synaptic cleft, leaving the synapse ready for additional transmission cycles.

As the chemical synapse requires the release and diffusion of neurotransmitter molecules across the synaptic cleft, this introduces a time delay every time an electrical signal needs to be transmitted from a cell to another. This is usually around 0.5 – 1 ms, but can be longer.

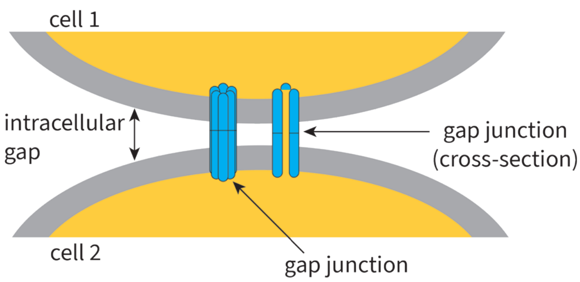


Figure 6.17. An electrical synapse. The cytoplasms of the two excitable cells are connected by pore-like structures called gap junctions.

5.2. Electrical synapses

Electrical synapses are rarer in our body than chemical synapses, and are employed when the activity of a large group of cells needs to be quickly coordinated (for example, the coordinated depolarization of cardiac muscle cells).

Unlike chemical synapses, there is no neurotransmitter present. Rather, two cells share pore-like structures called *gap junctions* that connect the cytoplasm of both cells (Figure 6.17). Each gap junction is made out of two half-channels named *connexons*. In turn, each connexon is assembled from 6 transmembrane proteins called *connexins*.

Through the gap junctions, ions and small molecules can pass from one side of the synapse to the other. In an electrical synapse, the intracellular gap is much smaller than the synaptic cleft of chemical synapses and the synaptic delay is minimal. Besides being faster than chemical synapses, electrical synapses are also *bidirectional*.

A comparison of chemical and electrical synapses is shown in Table 6.2.

6. BIOEXCITABILITY

We have seen that *excitable cells* are capable of responding to external stimuli by generating APs. We can evaluate how sensitive a cell is to a particular stimulus by employing two quantities called *rheobase* and *chronaxie*. We define the *rheobase* as the lowest intensity of an infinitely long

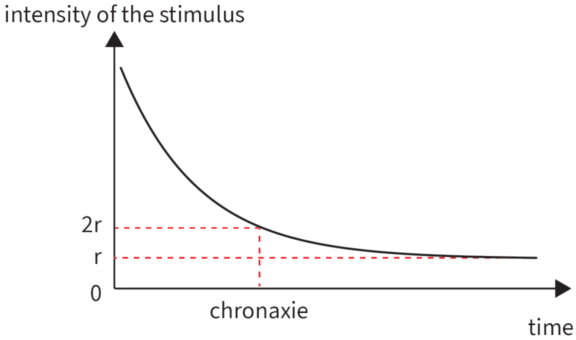


Figure 6.18. The relation between the intensity and the duration of a stimulus capable of eliciting a response in the excitable cell (Weiss' law). r = rheobase.

stimulus that is capable of triggering a response (an AP) in the excitable cell. The *chronaxie* is the time that you need to apply a stimulus that has double the intensity of the rheobase in order to trigger a response.

Figure 6.18 graphically shows the relationship between the intensity and duration of a stimulus that can generate a biological response. This is drawn according to an empirical (experimental) equation called Weiss' law:

$$\mathrm{i}\,(\mathrm{t}) = \frac{a}{t} + b \qquad (6.11)$$

where i(t) is the intensity of the stimulus, while *a* and *b* are positive constants.

Analyzing Weiss' law, we can see that:

- if $t \to \infty$, $i = b = r$ (the rheobase);
- if $i = 2b$, then $t = \frac{a}{b}$ = the chronaxie.

Accurate determination of the strength-duration (Weiss) curve for each patient is critical in setting the pacing efficiency of cardiac pacemakers.

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell 4th Edition: International Student Edition*. New York: Garland Science.

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Boron, W. F., & Boulpaep, E. L. (2017). *Medical Physiology* (3 ed.). Philadelphia: Elsevier.

Dillon, P. F. (2012). *Biophysics: A Physiological Approach*. Cambridge: Cambridge University Press.

Guyton, A. C., & Hall, J. E. (2005). *Textbook of Medical Physiology. Eleventh Edition*. Philadelphia: Elsevier.

Kielian, M. (2020). Enhancing host cell infection by SARS-CoV-2. *Science, 370*(6518), 765-766. doi:10.1126/science.abf0732

Nelson, D. L., Cox, M. M., & Hoskins, A. A. (2021). *Lehninger Principles of Biochemistry Eighth Edition*. New York: MacMillan Learning.

Sperelakis, N. (2012). Chapter 9 - Origin of Resting Membrane Potentials. In N. Sperelakis (Ed.), *Cell Physiology Source Book (Fourth Edition)* (pp. 121-145). San Diego: Academic Press.

Sperelakis, N. (2012). Chapter 10 - Gibbs–Donnan Equilibrium Potentials. In N. Sperelakis (Ed.), *Cell Physiology Source Book (Fourth Edition)* (pp. 147-151). San Diego: Academic Press.

Wright, S. H. (2004). Generation of resting membrane potential. *Adv Physiol Educ, 28*(1-4), 139-142. doi:10.1152/advan.00029.2004

Yartsev, A. Deranged Physiology. Retrieved from https://derangedphysiology.com/main/home

CHAPTER 7

MUSCLE CONTRACTION

Prerequisite knowledge

- Membrane resting potential
- Action potential
- Propagation of the action potential. Synapses

1. MOLECULAR MOTORS

1.1. Examples of molecular motors

Movement is a general characteristic of life. Organisms need to move in order to respond to different stimuli, to find food sources or breed. Internally, movement of nutrients or cell components is required in order to ensure survival. While some of these movements can be done passively (the organism is carried somewhere by water or wind, or the cellular component is carried by diffusion), there is also a need for active movement. This can be realized by *molecular motors*, cellular components that are capable of converting the chemical energy stored by the organism (many times, in molecules such as ATP) into mechanical energy.

Several types of molecular motors are present in the human body. Besides the actin-myosin motor, that will be the main focus of this lecture, we will mention a few more. For instance, the F_0F_1 ATP synthase in the mitochondrial membrane uses a rotary mechanism in order to synthesize ATP from ADP by using the transmembrane proton gradient as an energy source. Kynesin and dynein are molecular motors that transport organelles inside the cell by "walking" step by step using microtubules as a veritable "road". Thus, kinesin transports organelles and vesicles towards the plus end of a microtubule, while dynein facilitates transport in the opposite direction, towards the minus end. In presynaptic neurons, kinesin transports vesicles containing neurotransmitter towards the synapse. In the postsynaptic neuron dendrite, kinesin molecules transport receptors that are then inserted in the membrane.

1.2. ATP as the energy currency of the body

Many molecular motors employ the hydrolysis of ATP as their main energy source. We have already shown the hydrolysis reaction of ATP in a previous chapter, but let us present it here again in Figure 7.1 for the sake of convenience.

If we need the answer of why a molecule such as ATP is used in our body, we simply need to remember what we learned in Thermodynamics. It is impossible for our body to "burn" energy sources such as glucose in a single reaction step, as that would release a too high amount of energy. ATP has the advantage of being able to store the energy released stepwise in the glucose oxidation pathways and to be easily transportable to where the cell needs it for expenditure.

Why is, then ATP chosen to store this energy and not another molecule and why is energy released when ATP is broken down? The structure of ATP is quite unstable due to the proximity of three charged phosphate groups that repel each other. The usual way these are explained in biology textbooks is that the phosphate – phosphate bonds are "high energy" bonds which release a high amount of energy when they are broken. This is not really correct.

As a general rule of chemistry, **energy is required when chemical bonds are broken** and

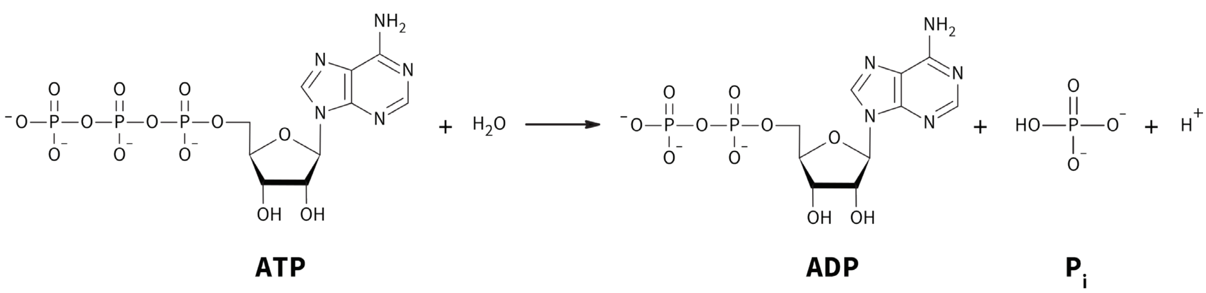


Figure 7.1. Hydrolysis of ATP to ADP and inorganic phosphate releases a high amount of energy (ΔG < 0).

energy is released when chemical bonds are formed. Thus, the simple breaking of a bond in ATP should, and does require energy. Where does the energy release come from then? Remember that, in the process of breaking ATP, ADP and inorganic phosphate (P_i) are formed at the same time. All these molecules will also have interactions with the solvent (water). **ADP and P_i are more stable energetically than ATP**, both from an enthalpic point of view (heat of formation), but also from an entropic point of view (entropy is higher for the ADP + P_i mixture that for ATP by itself). Remembering that the Gibbs free energy of a chemical reaction involves both an enthalpic and an entropic term, we can conclude that the Gibbs free energy of the products (ADP + P_i) is lower than that of ATP ($\Delta G < 0$). Energy is, thus, released in this reaction.

Our body uses a high amount of ATP in one day: the most commonly quoted figure is around our body's weight. However, at any given time only a tiny fraction of that amount (the usually quoted number is ~100 g) is present in the body. This is because ATP is continuously being used up (hydrolyzed) and then regenerated, mainly by the ATP synthase in mitochondria.

2. STRUCTURAL FEATURES OF SKELETAL MUSCLES

2.1. Muscle tissue in the human body

The human body contains three types of muscle tissues:

- **Skeletal muscle**, which controls the movement of the skeleton;
- **Smooth muscle**, lining the walls of hollow internal organs (viscera);
- **Cardiac muscle**, a specialized type of muscle tissue responsible for the contractions of the heart.

This chapter concerns itself mainly with the contraction of the skeletal muscle, which we will describe in detail. At the end, we will also mention the key differences between the contraction of the three muscle types.

2.2. Structure of skeletal muscles

The skeletal muscle is also called striated[1] muscle due to the appearance of skeletal muscle cells (called *muscle fibers* or *myocytes*) when observed under a microscope.

Let us briefly describe the structure of the skeletal muscle (Figure 7.2). Each individual muscle fiber is covered by a connective tissue layer called *endomysium*. Several muscle fibers are bundled into *fascicles*, and each fascicle, in turn, is covered by connective tissue called *perimysium*. Finally, fascicles are bundled together to form the muscle by connective tissue called *epimysium*. Towards the end of the muscle, the epimysium thickens and merges with a tendon or aponeurosis, which then attach the skeletal muscle to a bone. For muscles that attach directly to the bone, the epimysium merges with the periosteum of the bone.

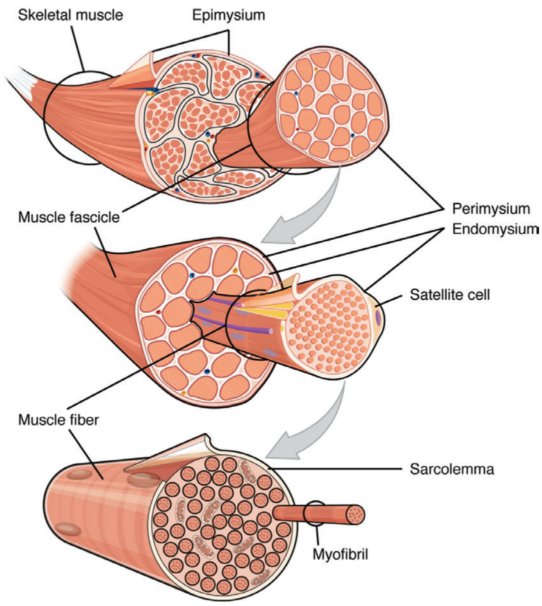


Figure 7.2. Structure of a skeletal muscle[2]. Detailed description is provided in the text.

2.3. Structure of the muscle fiber

Muscle fibers (Figure 7.3) are elongated, multi-nuclear cells. In humans, most muscle fibers are as long as the entire muscle (from tendon to tendon), therefore, their length can go up to tens of centimeters.

The plasma membrane of the muscle fiber is called the *sarcolemma*. Unlike the membrane of a typical cell, the sarcolemma has a large number of invaginations, called *T tubules* or *transverse tubules*. These are closed at the end, but enter deep inside the cell, allowing action potentials that propagate through the sarcolemma to be received towards the interior of the cell. The endoplasmic reticulum of the muscle fiber is called the *sarcoplasmic reticulum* (SR). The T tubules come into close proximity to the sarcoplasmic reticulum. The

1 A synonym for "striated" is "striped".

2 Image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). *Anatomy and Physiology 2e*. Retrieved from *https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction*

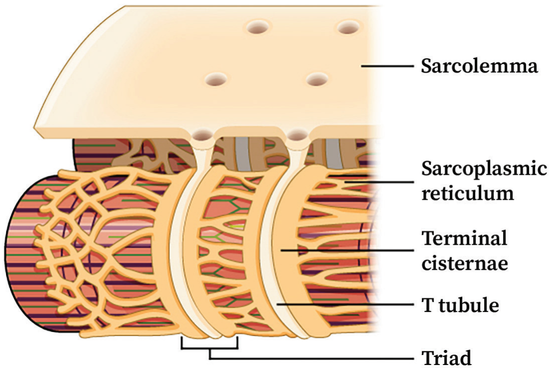


Figure 7.3. Triads of the skeletal muscle fiber[3]. Description is provided in the text.

structure formed by a T tubule with two *terminal cisternae* of the sarcoplasmic reticulum is called a *triad* (Figure 7.3). The T tubules and the terminal cisternae facilitate the transduction of the nervous impulses that depolarize the muscle fiber into muscle contraction, as we will see further. The sarcoplasmic reticulum acts as a reservoir for Ca^{2+} ions in the muscle fiber: Ca^{2+} is stored in the terminal cisternae at a concentration of ~400 µM. By comparison, the cytoplasmic (*sarcoplasmic*) concentration of Ca^{2+} in the relaxed muscle is kept very low (~0.05 – 0.1 µM).

Muscle fibers contain rod-like, contractile organelles called *myofibrils*. These have no membrane. Myofibrils are about 1 – 2 µm in diameter and extend the entire length of the muscle fiber. They consist of repeating subunits called *sarcomeres*. It is the myofibrils that give skeletal muscle fibers their striated aspect, due to the repeating sarcomeres. Each sarcomere (Figure 7.4) has a resting length of around 1.5 to 3.5 µm (on average ~2.2 µm).

The **sarcomere** (Figure 7.4) is the basic functional unit of the skeletal muscle. A sarcomere is delimited on either side by a disk called the *Z line*[4] (*Z disk*). Two types of protein filaments are attached to the Z disk:

- **thin filaments**, made up of three types of protein: *F-actin*, *tropomyosin* and *troponin*;
- **thick filaments**, which are made up of the protein myosin and attach to the Z disk via another protein called *titin*.

Microscope imaging of the muscle fiber allows the visualization of differently shaded bands inside the sarcomere (Figure 7.4, bottom panel). Besides the Z disks, we can observe:

- the I (*isotropic*) band extends on either side of the Z disks and contains actin filaments as well as the titin filaments binding myosin to the Z disk;

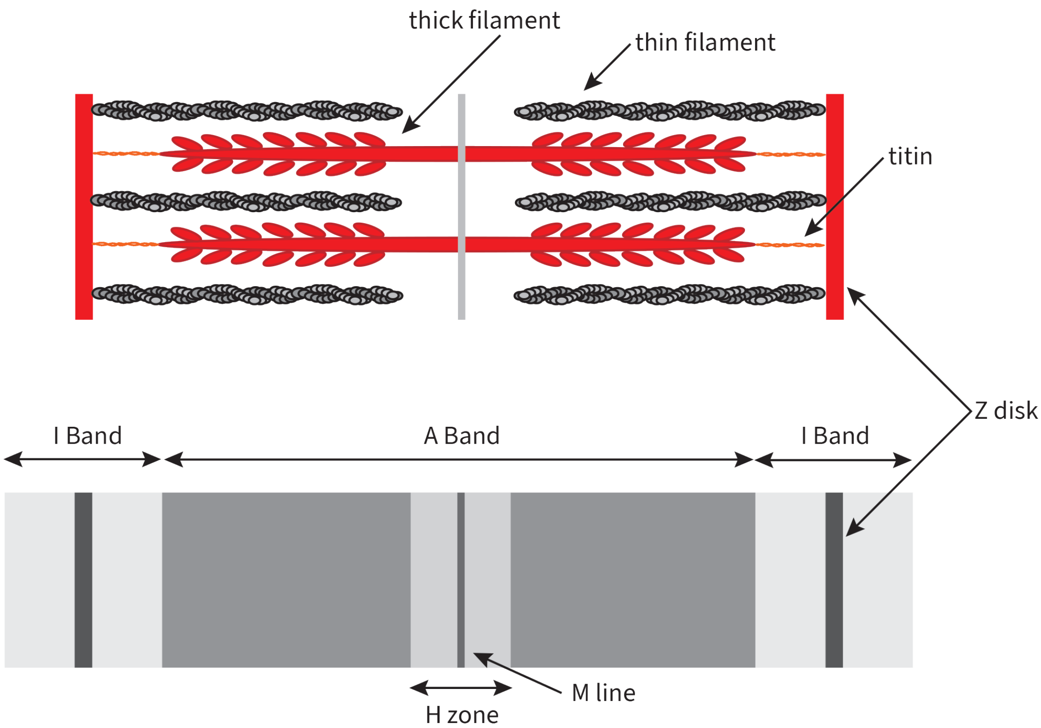


Figure 7.4. Structure of a sarcomere. Top panel: cartoon drawing of the sarcomere. Description is provided in the text. Bottom panel: When viewed using microscopy, the sarcomere can be observed as differently shaded bands.

3 Image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). *Anatomy and Physiology 2e*. Retrieved from *https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction*

4 From *zwischen* in German, which means “between”.

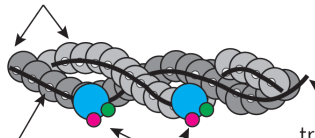


Figure 7.5. Composition of a thin filament. For better visualization, the two intertwining actin chains in the figure are shown with slightly different shades of grey (though they are functionally identical). Description is in the text.

▸ the A (*anisotropic*) band extends towards the middle of the sarcomere. It contains regions where myosin filaments extend, either overlapping or not with the thin filaments;
▸ towards the middle of the A band (middle of the sarcomere), a region can be observed called the H zone[5], where only myosin filaments exist;
▸ the middle of the H zone contains the M line[6], a disk-like transverse structure made up of a protein called *myomesin* that links and arranges the thick filaments.

2.4. The thin filaments

The thin filaments (also called "actin filaments") have a diameter of 6 – 10 nm. A thin filament (Figure 7.5) is formed by two intertwining chains of a protein called *F-actin* (we will call it just actin from now on), which is a polymer of globular actin (G-actin). Each G-actin subunit contains a binding site to which the head of a myosin protein can attach. When the muscle is in the relaxed state, the myosin binding sites of actin are covered by another filamentous protein called *tropomyosin*.

In addition to actin and tropomyosin, thin filaments also contain a complex of proteins called the *troponin complex*. The troponin complex is made up of three proteins: troponin C, troponin I and troponin T. Troponins are regulatory proteins: they can bind Ca^{2+} ions and, upon doing so, change their conformation, moving the tropomyosin filaments and exposing the myosin binding sites on actin. By controlling access to the myosin binding sites, troponins regulate muscle contraction, which is triggered when Ca^{2+} is released from the SR, as we will see further. Troponins are only found in the skeletal and cardiac muscle, but not in the smooth muscle.

2.5. The thick filaments

The thick filaments (also called "myosin filaments") have a diameter of ~ 15 nm and are solely composed of the protein *myosin*. They attach to the Z disks via a fibrillar protein called titin, as mentioned previously.

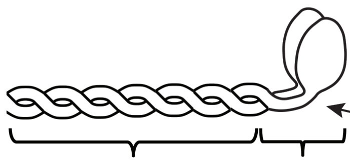


Figure 7.6. Myosin. HMM = heavy meromyosin; LMM = light meromyosin.

Myosin (Figure 7.6) is composed of two major subunits that can be separated when the protein is enzymatically cleaved:
▸ *heavy meromyosin* (HMM), which contains two domains: S-1 (the myosin head that can bind to actin) and S-2, that connects HMM to the second subunit, light meromyosin, forming a "neck" region;
▸ *light meromyosin* (LMM), a long, straight filament, essentially the "tail" of the myosin molecule.

Each thick filament contains hundreds of myosin molecules; the thick filaments are assembled in such a way (see Figure 7.4) that the heads are pointing towards the Z disks (away from the M line). The S-1 region of HMM (the head), besides being capable of binding actin, also has ATP-ase activity in the presence of Mg^{2+}. This activity increases by more than 200 times when the myosin head binds to actin. The ATP-ase activity of the myosin head causes a conformational change in the neck region that is essential for contraction, as we will see further.

3. THE MECHANISM OF MUSCLE CONTRACTION

3.1. Events leading to muscle contraction

As we know from a previous chapter, muscle cells are *excitable cells*: they are able to respond to external stimuli by generating action potentials (APs). For the skeletal muscle, the trigger for generating an action potential is always external, coming from a motor neuron (unlike the cardiac muscle, as we will see later). Skeletal muscles form synapses with motor neurons called *neuromuscular junctions*. A neuromuscular junction is an excitatory chemical synapse which uses acetylcholine (ACh) as a neurotransmitter (Figure 7.7).

A single motor neuron can innervate more than one muscle fiber. We say that a motor neuron together with all of the muscle fibers that it

5 From *heller* in German, which means "brighter".
6 From *Mittel* in German, which means "middle".

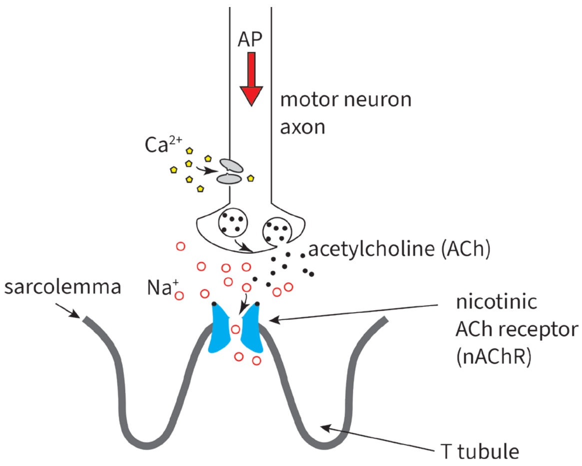


Figure 7.7. The neuromuscular junction. Detailed description is provided in the text.

innervates form a *motor unit*. For muscles involved in fine movements, only a few muscle fibers are innervated by one motor neuron. In muscles that are involved in powerful movements, up to one thousand muscle fibers might be innervated by a single motor neuron.

When an AP arrives through the axon of the motor neuron at the neuromuscular junction, voltage-gated Ca^{2+} channels in the axon membrane open, allowing Ca^{2+} to enter the neuron and triggering the fusion of vesicles containing ACh with the membrane of terminal buttons. ACh is released into the synaptic cleft of the neuromuscular junction, and will diffuse towards the membrane of the muscle fiber. There, ACh can bind to a specialized receptor called the nicotinic acetylcholine receptor (nAChR). nAChR is essentially a ligand-gated ion channel that will open and allow cations (Na^+ and K^+, sometimes Ca^{2+}) to pass through the membrane. Na^+ will enter the muscle fiber through the nAChR and depolarize the membrane. If the firing threshold of the muscle fiber membrane is reached, this will trigger an AP in the muscle fiber.

If you find it hard to understand why opening a channel (nAChR) that is permeable for both Na^+ and K^+ causes the depolarization of the membrane, remember that the membrane potential when more than one species is permeable is calculated by the Goldman equation which includes the permeabilities of the ions. K^+ is very permeable anyway through K^+ leak channels. Opening additional pathways for K^+ to exit the cell will only increase its permeability by comparatively little, while for Na^+ it's a very big effect, going from almost non-permeable to very permeable.

The excitation of the muscle fiber is followed by an excitation-contraction coupling event, which takes place at the triads. Triads are located at the junction between the A and I bands of the sarcomere. Therefore, depolarization of the sarcolemma will quickly result in consequences for the sarcomeres. Excitation-contraction coupling is achieved through the release of Ca^{2+} from the cisternae of the SR. This occurs as follows: when the sarcolemma is depolarized, Ca^{2+} channels in the membrane of the T tubules called dihydropyridine receptors (DHPRs) will activate. The DHPRs in the T tubule membrane physically interact with Ca^{2+} channels in the SR membrane called ryanodine receptors (RyRs). Thus, activation of the DHPRs will cause RyRs to open (Figure 7.8), allowing Ca^{2+} to exit the sarcoplasmic reticulum and increasing the cytoplasmic concentration of Ca^{2+}.

While both DHPRs and RyRs are Ca^{2+} channels, the main source of Ca^{2+} in skeletal muscle contraction results from the release of Ca^{2+} from the SR through the RyRs. The role of DHPR is mainly as a mediator that controls the opening of RyRs.

The release of Ca^{2+} from the SR connects the electrical stage to the mechanical stage of muscle contraction. When enough Ca^{2+} is available in the muscle fiber cytoplasm, this will cause binding of Ca^{2+} to troponin. As we previously mentioned, binding of Ca^{2+} to troponin causes a conformational change that moves the tropomyosin chains, exposing the myosin binding sites on actin. Thus, myosin will be free to bind actin, and the *crossbridge cycle* can begin.

3.2. The crossbridge cycle

The **crossbridge cycle** is the model that describes one cycle of contraction, in which a myosin head binds to actin and pulls the actin filament towards the center of the sarcomere. As we will see a little bit later, this results in the thin and thick filaments sliding along each other, thereby shortening the

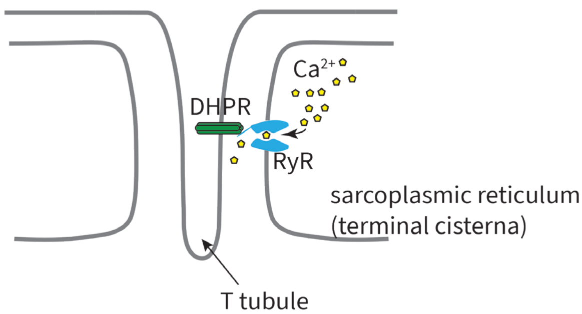


Figure 7.8. Release of Ca^{2+} from the SR. A triad (one T tubule + 2 terminal cisternae) is shown. For simplicity, only release of Ca^{2+} from the right cisterna is depicted. An action potential that depolarizes the T tubule membrane will activate DHPRs, which, in turn, will open Ca^{2+} channels (RyRs) in the sarcoplasmic reticulum membrane, causing Ca^{2+} release from the SR cisternae into the cytoplasm of the muscle fiber.

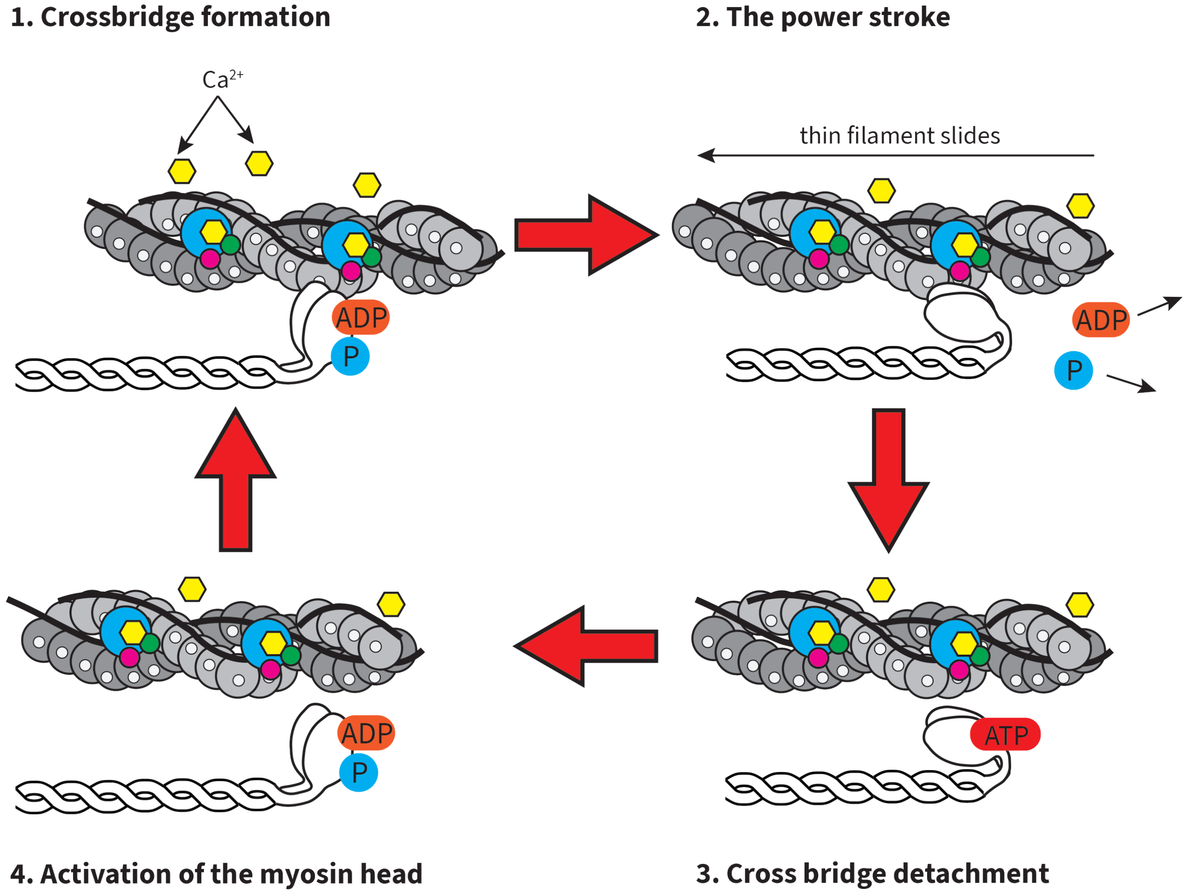


Figure 7.9. The crossbridge cycle. The cycle can begin once Ca^{2+} attaches to troponin and the myosin binding sites on actin are uncovered by the movement of myosin. Detailed description of each step in the cycle is given in the text. P = inorganic phosphate.

sarcomere.

There are four main steps to the crossbridge cycle, which we will describe in turn (Figure 7.9):

- *Step 1. Crossbridge formation*. Once the myosin binding sites on actin are revealed by the movement of tropomyosin, a myosin head which is in the active, or "cocked" position, binds to one such binding site. A crossbridge is formed between the thick and the thin filaments involved;
- *Step 2. The power stroke*. Release of ADP and P_i causes the neck of the myosin molecule to bend, pulling on the thin filament as it does so. As a result, the thin filament is pulled towards the center of the sarcomere (towards the M line) and the sarcomere shortens;
- *Step 3. Crossbridge detachment*. If ATP is present, it can attach to the myosin head, causing detachment of the myosin head from its binding site on actin;
- *Step 4. Activation of the myosin head*. ATP is hydrolyzed by myosin, providing enough energy to switch the myosin head back to the active, or "cocked" position. The cycle is ready to restart.

As you can see from the above description, ATP is required for both the formation and the breaking of crossbridges. In the absence of ATP, muscles remain in a permanently contracted state called *rigor mortis*. Rigor mortis can set in a few hours after death (when cellular sources of ATP are depleted), and lasts for several more hours, disappearing after the muscle fibers start to decompose.

3.3. The sliding filament model of contraction

As a result of the crossbridge cycle, the thin filaments are pulled towards the center of the sarcomere (Figure 7.10). This results in the thin filaments sliding along the thick filaments and the Z disks moving towards the M line. Note that the length of an individual myosin filament or that of an actin filament does not change, but the sarcomere shortens as the Z disks are pulled closer to each other.

During contraction, the I bands shorten and eventually disappear, while the length of the A bands (where the myosin filaments are present) stays constant. In one crossbridge cycle, a sarcomere is shortened by approximately 10 nm. Note that even in a single short contraction (twitch), the crossbridge cycle will repeat itself several times.

As the muscle fiber is isovolumetric (it keeps the same volume), shortening will cause a radial bulge of the muscle fiber. You can easily observe

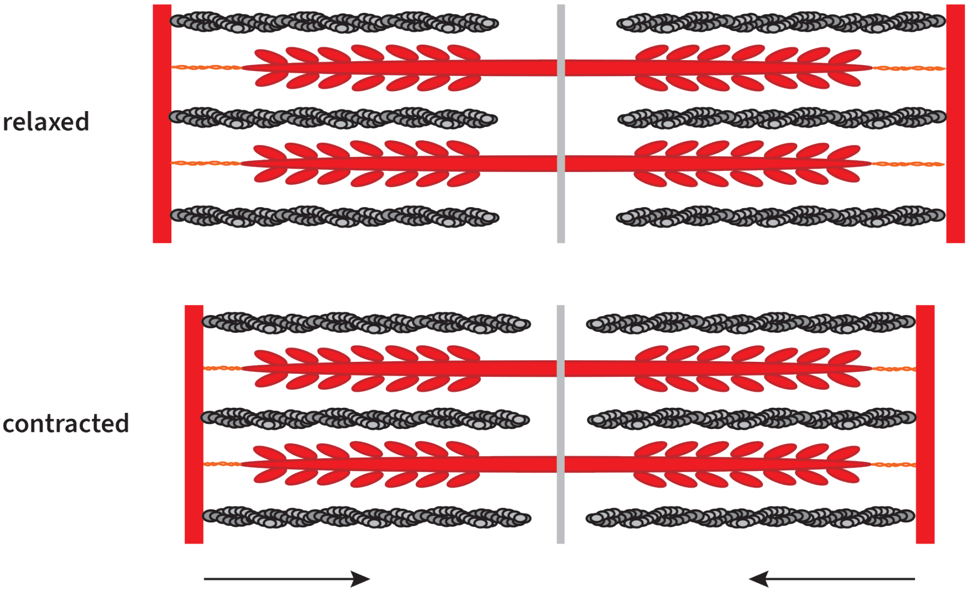


Figure 7.10. Contraction of the sarcomere. The myosin heads pull on the actin filaments, drawing the Z disks towards the M line.

this bulging while contracting your biceps.

The crossbridge cycles will keep on running so long as Ca^{2+} and ATP are present in the cytoplasm of the muscle fiber. The rate of sarcomere shortening varies depending on the type of muscle fiber. When the muscle fiber stops being excited by the motor neuron, relaxation of the muscle fiber will occur. The RyR channels in the sarcoplasmic reticulum membrane close and Ca^{2+} ions are returned to the sarcoplasmic reticulum through the action of a pump called the *sarcoplasmic/endoplasmic reticulum Ca^{2+}-ATPase* (SERCA). When not enough Ca^{2+} is available in the cytoplasm of the muscle fiber, tropomyosin returns to blocking the myosin binding sites; the actin filaments are free to slide back towards their relaxed position and the sarcomere returns to its relaxed length.

3.4. Types of muscle fibers

As we have seen, ATP is essential for the correct function of the muscle fibers, in both contraction and relaxation. Three main sources of ATP regeneration are used in muscle fibers, which you will study in more detail in Biochemistry:

- *The phosphagen system* (ATP – creatine phosphate) uses a chemical called creatine phosphate (CP) in order to regenerate ATP from ADP, in the presence of an enzyme called creatine kinase;
- *The aerobic metabolism*, which comprises pathways that produce high amounts of ATP in the presence of oxygen;
- *Anaerobic glycolysis*, producing comparatively smaller amounts of ATP in the absence of oxygen.

Depending on their contraction speed and method of ATP regeneration, muscle fibers can be classified into several types. The main types are (Table 7.1): slow oxidative (type I), fast oxidative (type IIa) and fast glycolytic (type IIx) fibers. Most muscles contain all three types of fibers, but in different proportions.

Slow oxidative fibers (Type I) contract slower than the other types and employ the aerobic metabolism as a method of regenerating ATP. They contain more mitochondria than glycolytic fibers and also contain *myoglobin*, an oxygen-binding protein similar to hemoglobin. Myoglobin gives these fibers a characteristic red color. These

Table 7.1. Characteristics of different types of muscle fibers.

	Slow-twitch, Type I (slow oxidative)	Fast-twitch, Type IIa (fast oxidative)	Fast-twitch, Type IIx (fast glycolitic)
Type of activities	Maintaining posture, endurance activities	Intermediate movements (e.g. walking)	Quick, powerful movements
Diameter of fiber	Small	Large	Large
Force developed	Low	High	Very high
Main energy source	Aerobic	Aerobic	Anaerobic
Contraction speed	Slow	Fast	Very fast
Resistance to fatigue	High	Low	Very low
No. of mitochondria	High	Medium	Low

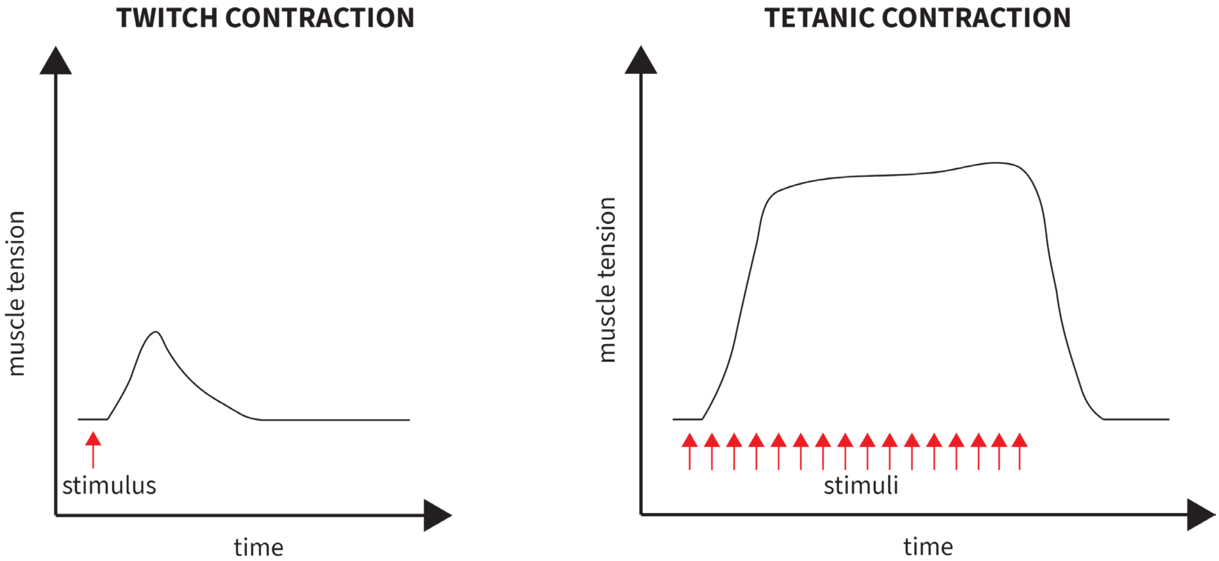


Figure 7.11. Twitch contraction (left) and tetanic contraction (right). Description is provided in the text.

fibers can be used for extensive amounts of time without fatigue. They are used mainly for maintaining posture.

Fast glycolytic fibers (Type IIx) use glycolysis as their main energy source and fatigue very quickly. They are used for rapid, powerful movements. These do not have myoglobin and have a white color.

Fast oxidative fibers (Type IIa) have intermediate properties between the other two types. They produce ATP more quickly than Type I fibers, but fatigue slower than Type IIx fibers. They are typically used for movements such as walking.

4. MECHANICS OF MUSCLE CONTRACTION

4.1. Types of muscle contraction

Through the formation of crossbridges, the muscle generates *tension* (we also simply say that the muscle contracts). While we previously described that the muscle shortens following contraction, this is not always true, and depends on the external forces that act upon the muscle. For example, when lifting an object, the muscle tension is the force exerted on the object by the contracting muscle and the external force is the force exerted on the muscle by the object (usually the object's weight). Whether a fiber shortens or not depends on the relative magnitudes of the two opposing forces. Depending on the external force and the tension developed in the muscle, the main types of muscle contraction are:

- *Isotonic* contractions, where a constant tension is maintained in the muscle during contraction. Lifting a weight from the ground is an example of an isotonic contraction. Isotonic contractions can be further classified as: *concentric* (resulting in the muscle shortening) and *eccentric* (resulting in the muscle lengthening). Note that even in eccentric contractions the crossbridge cycle still occurs even though the sarcomere is lengthening, but the external force is larger than the tension developed by the muscle;
- *Isometric* contractions, where tension is generated without changing the length of the muscle. Pushing against a wall is an example of an isometric contraction.

A single action potential (single stimulus) initiated by a motor neuron will cause a single contraction, termed *twitch contraction*. If the frequency of stimulation is very high, the twitch contractions can become closer and closer together, until they are virtually summed up, resulting in a permanent state of contraction called *tetanic contraction* (Figure 7.11). Tetanic contractions occur normally: e.g. when holding up a heavy weight.

Sometimes the terms “tetanus” or “physiological tetanus” are used as synonyms for “tetanic contraction”. We preferred not using “tetanus” in this text, as it might be mistaken for the disease called by the same name. Tetanus (lockjaw) is a disease caused by the bacterium *Clostridium tetani* characterized by severe muscle spasms. This disease can be fatal, but is prevented by vaccination.

Another word that sounds similar is “tetany”. Tetany is a symptom that involves involuntary muscle contractions, usually caused by electrolyte imbalance.

Our muscles generate a small amount of tension most of the time. This is achieved by the contraction of only a few of the muscle fibers in a muscle and results in a state of partial contraction called *muscle tone*, which helps maintain posture.

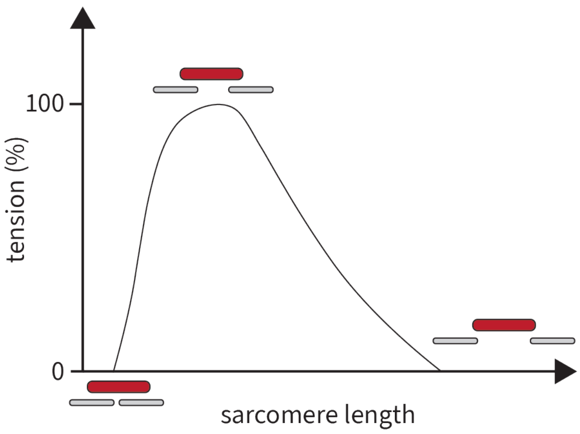


Figure 7.12. Length – tension relationship. The overlap of the thin filaments (grey bars) and the thick filaments (red bars) is shown in three different positions. Note that the optimal length varies depending on the type of muscle.

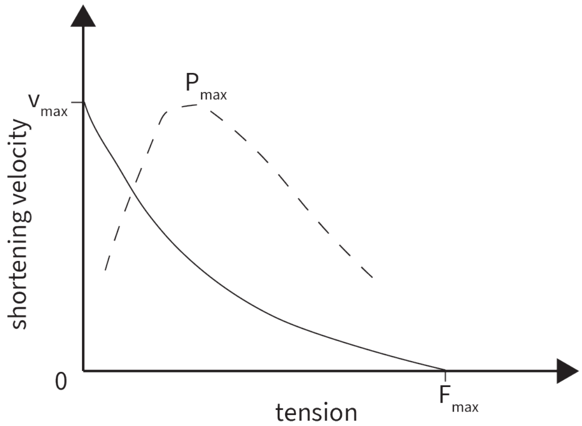


Figure 7.13. Tension – shortening velocity relationship. The solid line shows the plot of the Hill equation (v *vs* F). The dashed line shows a plot of the developed power in the muscle, with a maximum occurring at ~ $0.3F_{max}$).

Muscle tone is reduced during sleep.

4.2. The length – tension relationship in muscle

The force (tension) that the muscle is able to generate depends on the length of the muscle (length of the sarcomeres, to be more precise). When the muscle is fully shortened or fully elongated, it is not able to generate tension, with the maximum tension being generated at an intermediate length of the muscle (Figure 7.12).

The variation seen in Figure 7.12 is explained by the degree of overlap between the thin filaments and thick filaments in the sarcomere, and thus the possible number of crossbridges that can be formed.

When the sarcomere is fully elongated, the thin and thick filaments barely overlap, and none or a very limited number of crossbridges can be formed. At intermediate length (~2.6 – 2.8 µm) of the sarcomere, there is an optimal overlap between the thin and thick filaments, a high number of crossbridges can be formed and, thus, the tension that can be developed is maximal. When the sarcomere is completely shortened, even though a large number of crossbridges can form, myosin cannot pull on the thin filaments any more as they have nowhere to go. Thus, no force is developed.

4.3. The tension – shortening velocity relationship

The relationship between muscle tension and the shortening velocity is described by Hill's equation. This is a hyperbolic dependence which can be written as:

$$(F + a)(v + b) = c \qquad (7.1)$$

where a, b, and c are constants that depend on the muscle, F is the developed tension and v is the shortening velocity.

It can be observed that, if $F = 0$, v has to increase to a maximal value, v_{max}. The same is true in reverse, if $v = 0$, $F = F_{max}$ (Figure 7.13).

From equation (7.1), the power developed by the muscle can also be described, knowing that:

$$P = F \cdot v \qquad (7.2)$$

At $F = F_{max}$ and $v = v_{max}$ the developed power will be 0, as either $v = 0$, or $F = 0$. The maximal power is obtained for intermediate values of F and v ($F \sim 0.3F_{max}$).

5. THERMODYNAMICS OF MUSCLE CONTRACTION

Let us give a brief consideration to the process of muscle contraction from a thermodynamic point of view. Muscles convert chemical energy (stored in ATP molecules) into mechanical work. However, according to the second law of thermodynamics, such a process also inevitably produces heat. In fact, skeletal muscle is the main thermogenetic organ in our body.

Heat is generated in the muscle when ATP is hydrolyzed by myosin (during contraction), but also in the relaxed state, through the activity of SERCA, which hydrolyzes ATP in order to sequester Ca^{2+} ions in the SR.

An example of a process where the muscles are used exclusively to produce heat, and not work, is shivering, repeated involuntary contractions that can rapidly increase the body's temperature.

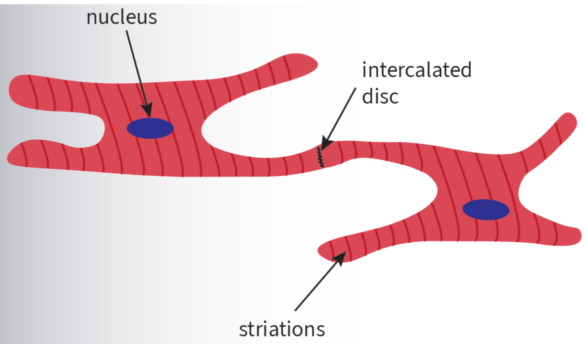


Figure 7.14. Two cardiomyocytes connected by an intercalated disc.

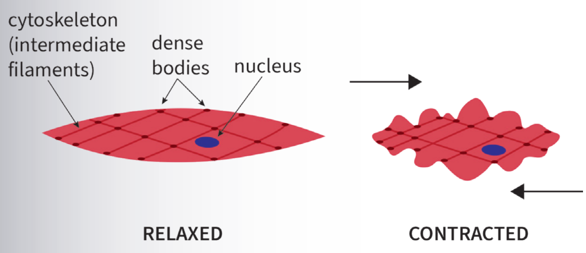


Figure 7.15. Contraction of a smooth muscle fiber.

6. SMOOTH AND CARDIAC MUSCLE

We will briefly list the main differences between striated muscle and the other two muscle types, cardiac and smooth muscle.

6.1. Cardiac muscle

Cardiac muscle is similar in structure with skeletal muscle, having thick and thin filaments, though a less developed sarcoplasmic reticulum. The cardiac muscle cells are called *cardiomyocytes*. Cardiomyocytes have the shape of branched cylinders (Figure 7.14) and are much shorter (~ 100 – 200 µm in length) than skeletal muscle fibers. Unlike skeletal muscle fibers, cardiomyocytes have usually only a single nucleus.

Cardiomyocytes are closely connected to one another through structures called *intercalated discs*. The intercalated discs contain structures that ensure the adhesion of cardiomyocytes to each other called *desmosomes* as well as electrical synapses (*gap junctions*).

As for the skeletal muscle, contraction of cardiomyocytes is also activated by an increase of intracellular Ca^{2+}, but the source of Ca^{2+} is different – most of it arrives in the cardiomyocyte from the extracellular medium, during an action potential. Thus, as extracellular Ca^{2+} enters the cell during the action potential, the shape of the action potential is different in cardiomyocytes compared to skeletal muscle fibers.

The source of the action potential in cardiac muscle is the heart tissue itself, not motor neurons, as for skeletal muscle: some cardiac muscle cells are pacemaker cells, that periodically generate their own action potentials. These are quickly propagated inside the heart through the gap junctions. Thus, the interconnected cardiomyocytes work as a unit ensuring coordinated contraction of the heart (they form a *functional syncytium*). More details regarding this will be given in the practical activities, when studying Electrocardiography (ECG).

6.2. Smooth muscle

Smooth muscle fibers are smaller, mononucleated and lack the highly repeating organization of the skeletal muscle fibers. Thus, there are also no myofibrils or sarcomeres present. Unlike striated muscle, smooth muscle is not under voluntary control. Thick and thin filaments still exist, but they are distributed throughout the cell. No troponin exists, and the regulatory role is taken up by a different Ca^{2+}-binding protein called calmodulin.

Similar to the cardiac muscle, the sarcoplasmic reticulum is less developed and the main source of Ca^{2+} during contraction is extracellular. The thin filaments are attached to protein structures called *dense bodies* that are, in turn, also attached to the cell membrane. When the smooth muscle fiber contracts, the sarcolemma is pulled towards the interior of the cell, shortening the fiber in a helical (corkscrew-like) movement (Figure 7.15).

REFERENCES

Adams, V. (2018). Electromyostimulation to fight atrophy and to build muscle: facts and numbers. *J Cachexia Sarcopenia Muscle, 9*(4), 631-634. doi:10.1002/jcsm.12332

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2012). *Biochemistry. Seventh Edition*. New York: Freeman and Company.

Biga, L. M., Bronson, S., Dawson, S., Harwell, A., Hopkins, R., Kaufmann, J., . . . Runyeon, J. (2019). *Anatomy & Physiology*. Retrieved from https://open.oregonstate.education/aandp/chapter/10-5-types-of-muscle-fibers/

Bojsen-Moller, J., & Magnusson, S. P. (2019). Mechanical properties, physiological behavior, and function of aponeurosis and tendon. *J Appl Physiol (1985), 126*(6), 1800-1807. doi:10.1152/

japplphysiol.00671.2018
Boron, W. F., & Boulpaep, E. L. (2017). *Medical Physiology* (3 ed.). Philadelphia: Elsevier.
Fan, R., & Lai, K. O. (2022). Understanding how kinesin motor proteins regulate postsynaptic function in neuron. *Febs j, 289*(8), 2128-2144. doi:10.1111/febs.16285
George, A. L., Jr. (2005). Inherited disorders of voltage-gated sodium channels. *J Clin Invest, 115*(8), 1990-1999. doi:10.1172/JCI25505
Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). *Anatomy and Physiology 2e*. Retrieved from https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction
Guyton, A. C., & Hall, J. E. (2005). *Textbook of Medical Physiology. Eleventh Edition*. Philadelphia: Elsevier.
Harris, A. J., Duxson, M. J., Butler, J. E., Hodges, P. W., Taylor, J. L., & Gandevia, S. C. (2005). Muscle fiber and motor unit behavior in the longest human skeletal muscle. *J Neurosci, 25*(37), 8528-8533. doi:10.1523/JNEUROSCI.0923-05.2005
Krans, J. L. (2010). The Sliding Filament Theory of Muscle Contraction. *Nature Education, 3*(9).
Kuo, I. Y., & Ehrlich, B. E. (2015). Signaling in muscle contraction. *Cold Spring Harb Perspect Biol, 7*(2), a006023. doi:10.1101/cshperspect.a006023
Lanner, J. T., Georgiou, D. K., Joshi, A. D., & Hamilton, S. L. (2010). Ryanodine receptors: structure, expression, molecular details, and function in calcium release. *Cold Spring Harb Perspect Biol, 2*(11), a003996. doi:10.1101/cshperspect.a003996
Nathan, R., Getz, W. M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D., & Smouse, P. E. (2008). A movement ecology paradigm for unifying organismal movement research. *Proc Natl Acad Sci U S A, 105*(49), 19052-19059. doi:10.1073/pnas.0800375105
Periasamy, M., Herrera, J. L., & Reis, F. C. G. (2017). Skeletal Muscle Thermogenesis and Its Role in Whole Body Energy Metabolism. *Diabetes Metab J, 41*(5), 327-336. doi:10.4093/dmj.2017.41.5.327
Seow, C. Y. (2013). Hill's equation of muscle performance and its hidden insight on molecular mechanisms. *J Gen Physiol, 142*(6), 561-573. doi:10.1085/jgp.201311107
Wada, M., Kuratani, M., & Kanzaki, K. (2013). Calcium kinetics of sarcoplasmic reticulum and muscle fatigue. *The Journal of Physical Fitness and Sports Medicine, 2*(2), 169-178.

CHAPTER 8

PHOTOBIOLOGY

Prerequisite knowledge

- Structure of the atom
- Waves and their characteristics

1. RADIATION

The emission or propagation of energy over a distance through the means of waves or particles is called *radiation*. Depending on its physical nature, we can classify radiation as *electromagnetic waves*, *acoustic waves*, *particle radiation*, etc.

Radiation can also be classified by its energy: high energy radiation is capable of extracting electrons from the outer electron shell of atoms and molecules. The process is called *ionization*. Radiation that can ionize atoms or molecules is called *ionizing radiation*. If the energy carried by the radiation is lower, such that it cannot ionize matter, we call it *non-ionizing radiation*.

Two chapters will present the interaction of radiation with biological systems:

- **Photobiology** (the current chapter) studies the effects of non-ionizing electromagnetic radiation on living organisms;
- **Radiobiology** (the following chapter) studies the effects of ionizing radiation (particle or electromagnetic) on living organisms.

Further interactions of radiation with living tissue are described in the chapters on Medical imaging and Physical factors in therapy.

2. ELECTROMAGNETIC RADIATION

2.1. Electromagnetic radiation as a wave

Electromagnetic (EM) radiation is the simultaneous propagation in space of two fields: an electric field and a magnetic field. In classical physics, EM radiation can be described as a wave (Figure 8.1). Note that physicists also use the term *light* interchangeably with EM radiation, not referring necessarily only to visible light. We will also occasionally use it with the same meaning in this text.

The two propagating fields (electric and magnetic) oscillate at 90° to both each other and to their direction of travel. Consequently, EM waves are classified as *transverse waves* (the opposite of this are *longitudinal waves* such as acoustic waves, which we study in a different chapter).

It is usually simpler to represent only the electric field when characterizing EM radiation. We can represent the oscillations of the electric field either as a function of the distance travelled by the EM radiation, or as a function of the time of travel (Figure 8.2).

From the two representations in Figure 8.2 we can characterize the EM wave by the following physical quantities:

- the **wavelength** of the EM radiation is the distance between two consecutive maxima. This is denoted with the Greek letter λ. As it is a distance, we measure wavelength in meters (m);
- the **period** of the EM radiation is the time that it takes for the radiation to go from one maximum

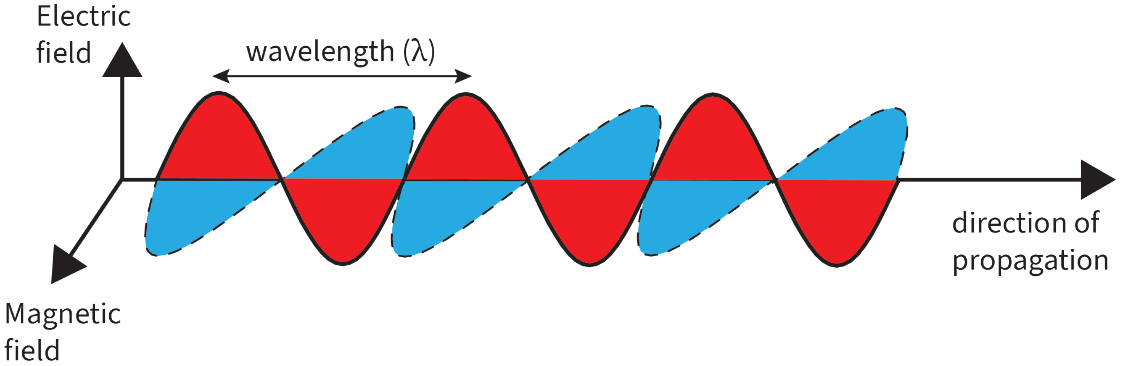


Figure 8.1. An electromagnetic wave is the simultaneous propagation of an electric (red) and a magnetic (blue) field. Their oscillations occur at 90° to both each other and to the direction of travel.

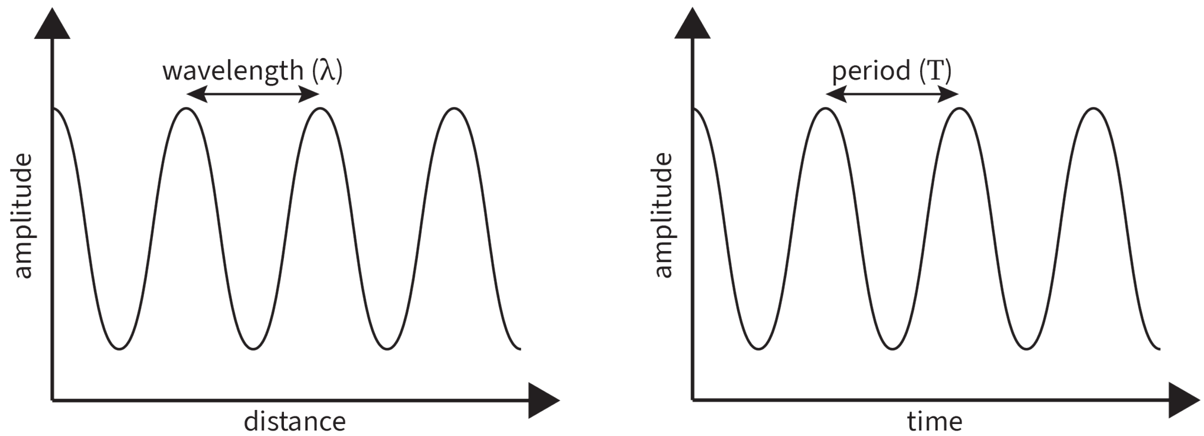


Figure 8.2. General representations of a wave. Represented are the travel of the wave across a distance (left panel) and the time course of the wave in a particular point (right panel).

to another. This is denoted by the letter T. As it is a time interval, we measure period in seconds (s).

In vacuum, electromagnetic radiation travels at a fixed speed, denoted as *c*, which is approximately equal to ~$3{\cdot}10^8$ m/s (~300000 km/s). We can easily see that:

$$c = \frac{\lambda}{T} \tag{8.1}$$

Thus, wavelength and period are directly proportional (a wave of high wavelength will have a high period). More commonly than the period, we prefer to use its inverse quantity, which is called the **frequency** of the wave (noted usually with the Greek letter ν):

$$\upsilon = \frac{1}{T} \tag{8.2}$$

The frequency essentially shows how many times in a given period of time the wave will go through maxima (how many oscillations occur in one unit of time). We measure frequency in s^{-1} which is called Hz (hertz). An EM wave with ν = 1 Hz will have one oscillation per second, an EM wave with ν = 100 Hz will have 100 oscillations per second, etc.

We can, then, write:

$$c = \lambda \cdot \nu \tag{8.3}$$

This is equivalent to:

$$\upsilon = \frac{c}{\lambda} \tag{8.4}$$

Equations (8.3) and (8.4) show that, for a given wave, **wavelength and frequency are inversely proportional** (when one is high, the other is low).

2.2. Electromagnetic radiation as particle radiation

There are properties of EM radiation that cannot be explained by its nature as a wave, such as the photoelectric effect. Therefore, in quantum mechanics, EM radiation is treated as being propagated as uncharged elementary particles of zero resting mass called *photons*. Photons always travel at a maximum speed, equal to *c*, no matter the medium through which they propagate.

We say that the energy of EM radiation is *quantized*[1], and, for a given monochromatic (of only one frequency) EM radiation, the energy of one photon can be calculated using the Planck-Einstein relation:

$$E = h\upsilon = \frac{hc}{\lambda} \tag{8.5}$$

where E is the energy of a photon (measured in J in the International System of Units), ν is the frequency of the radiation and h is Planck's constant ($h = 6.62 \cdot 10^{-34}$ J·s).

The energy of an individual photon is extremely small compared to that of macroscopic objects. Therefore, instead of using J (joules) as a unit of measurement for the energy of photons and subatomic particles it is usually preferred to use another, much smaller unit, called the electron-volt (eV).

1 eV is the work required to move one electron through a potential difference of 1 V. Note that eV is not a submultiple of the volt (V)! The two measure two completely distinct quantities: V measures electric potential, while the eV measures energy.

In J, the value of 1 eV is:

$$1\ eV = 1.602 \cdot 10^{-19}\ J \tag{8.6}$$

We say that EM radiation has a *dual nature* as both a particle and a wave: some of its properties can be explained by light being a wave (reflection,

1 *Quantized* = it can only have certain (discrete) values.

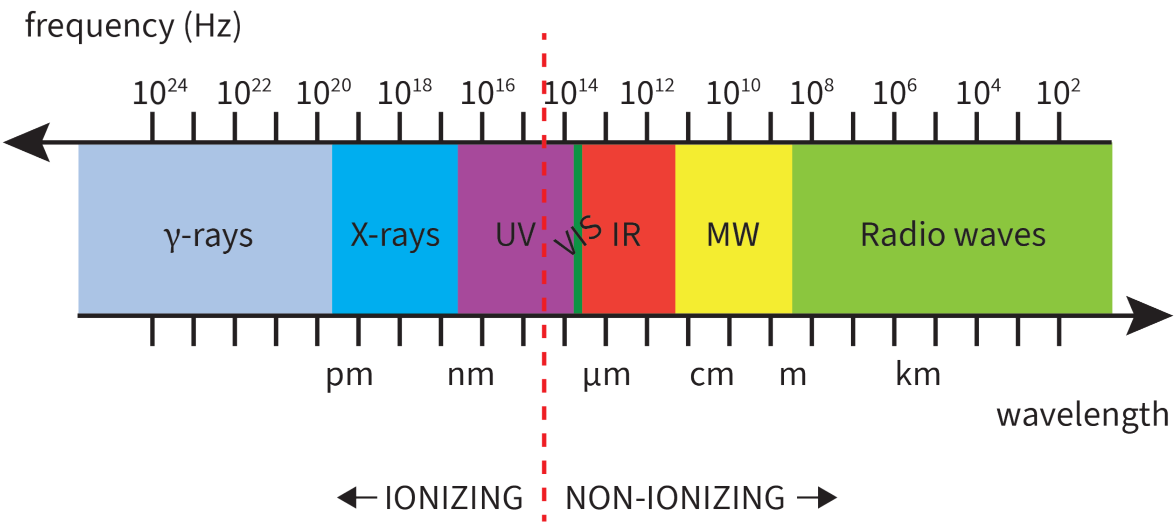


Figure 8.3. The electromagnetic spectrum. Wavelengths increase from left to right. The scales for both frequency and wavelength are logarithmic. Domains are colored arbitrarily for ease of reading. EM radiation can be classified as ionizing (γ-rays, X-rays, part of the UV domain) or non-ionizing (the rest of the EM spectrum). Details will be given further in the text. UV = ultraviolet; VIS = visible light; IR = infrared; MW = microwaves.

refraction, diffraction, etc.), while other properties (the photoelectric effect, absorption and emission) can be explained by light being carried by particles (photons).

2.3. The EM spectrum

Summing up, we can conclude that a given monochromatic radiation can be equally characterized by one of the following: its **wavelength**, its **period**, its **frequency** or the **energy** of its photons, as all of these are related to each other by constants.

Depending on their frequency (or energy, or wavelength, etc.) we can sort the EM radiation into different domains. We call the totality of frequencies (or energies, or wavelengths, etc.) of EM radiation the *electromagnetic spectrum* (Figure 8.3). As frequencies of EM radiation can take essentially any value, this is a continuous spectrum.

In the representation of Figure 8.3, wavelength increases from left to right, and equally frequency and energy decrease from left to right. We can distinguish, thus, the following domains in the EM spectrum:

- **γ rays** (or γ radiation) are the highest energy type of radiation. These are generally produced by the radioactive decay of the nuclei of atoms and have wavelengths smaller than 10 pm (10^{-11} m);
- **X-rays**, situated between γ rays and UV radiation with wavelengths of ~10 pm – 10 nm (~10^{-11} – 10^{-8} m). They are produced via transitions in the electron cloud of atoms (see the Medical imaging chapter for a description of the process);
- **ultraviolet (UV) radiation**, at shorter wavelengths than the visible spectrum, with wavelengths of 10 nm – 400 nm (10^{-8} m – $4\cdot10^{-7}$ m);
- **visible light** is the part of the EM spectrum that we can directly perceive using our eyes. It represents a comparatively tiny part of the EM spectrum, as you can see from Figure 8.3! Visible light is generally considered to be the portion of the spectrum between **400 – 760 nm** ($4\cdot10^{-7}$ m – $7.6\cdot10^{-7}$ m), though the limits for each individual person differ. In a given medium, the color of visible light is given by its frequency. For example, if red light passes from air to water, it changes its wavelength, but it is still perceived as red. The visible light spectrum is shown in better detail in Figure 8.4;
- **infrared (IR) radiation**, with wavelengths between 760 nm to 1 mm ($7.6 \cdot 10^{-7}$ m – 10^{-3} m). All objects at temperatures above 0 K, including the human body, emit IR radiation;
- **microwaves (MW)**, with wavelengths of approximately 1 mm to 30 cm (10^{-3} m – $3\cdot10^{-2}$ m);

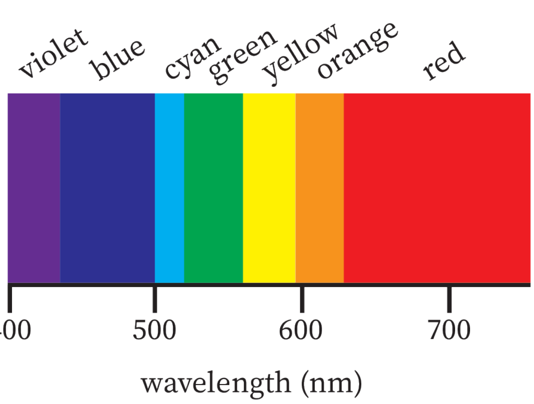


Figure 8.4. Colors of the visible spectrum. The wavelength corresponds for propagation of light in air. Note that this is not a "true" representation of color at every wavelength, as different shades of the base color are perceived depending on wavelength and luminance. The exact process of color perception is described in the chapter on the Biophysics of vision.

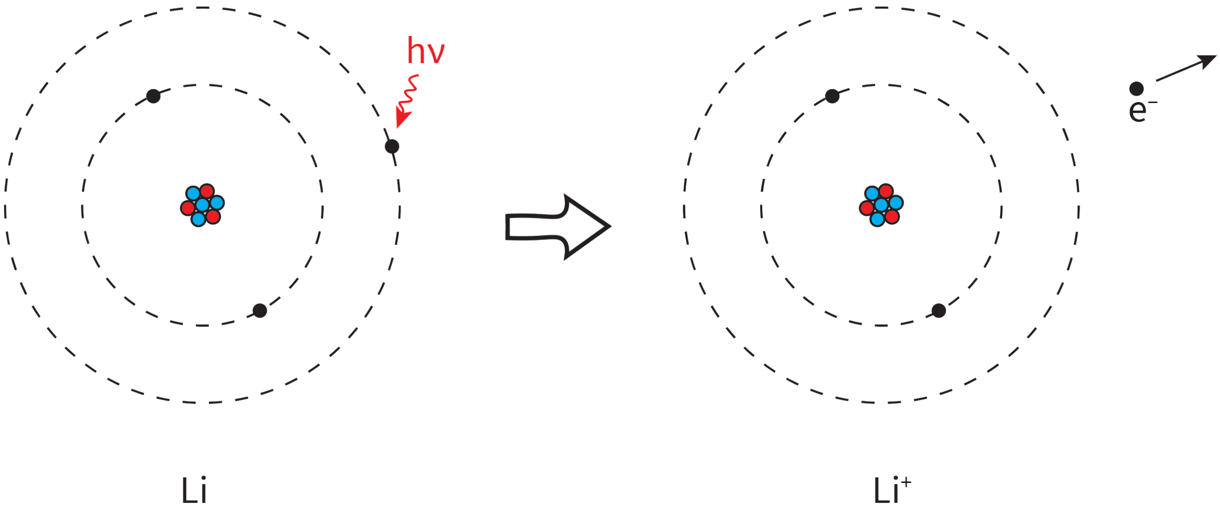


Figure 8.5. Ionization of the Li atom. The Li nucleus has 3 protons (red circles) and 4 neutrons (blue circles). Li has 3 electrons (small, black circles) distributed in two atomic orbitals (two energy levels). The outer electron can be removed from the Li atom if enough energy is provided by incoming radiation, forming the Li^+ ion and a free electron (e^-).

- **radio waves**, with wavelengths above 30 cm, that can go up to tens or hundreds of km (there is no known upper limit).

For the sake of simplicity, we gave clear limits of energy between γ and X photons in this text as well as in Figure 8.3. However, there is actually a high overlap between the energies of these two domains. The actual difference between what makes a photon into a γ or an X photon is its origin: we call high energy photons that result from processes inside the nucleus γ photons, while high-energy photons that result from processes outside the nucleus are called X photons.

2.4. Ionizing and non-ionizing EM radiation

As we said in the beginning, we can classify radiation depending on whether it has the capability of ionizing atoms or molecules (Figure 8.5). We consider the limit of energy between non-ionizing and ionizing energies as **10 eV**. Thus, radiation with energies of 10 eV and above is ionizing, and that with energies below 10 eV is non-ionizing.

For EM radiation, the limit of 10 eV corresponds to a wavelength of ~ 124 nm, which is found in the UV domain (Figure 8.3). Thus, we conclude that, for the radiation in the EM spectrum:

- **γ rays, X-rays** and **UV radiation with λ ≤ 124 nm** are **ionizing** radiation;
- **UV radiation with λ > 124 nm, visible light, infrared radiation**, **microwaves** and **radio waves** are **non-ionizing** radiation.

The process of ionization is damaging to matter, and, by extension, to living organisms as ionization of biologically important molecules (DNA, proteins, lipids, etc.) can alter their function and damage the cellular mechanisms. This will be discussed in more detail in the following chapter.

Even though, conventionally, a limit of 10 eV is chosen to separate ionizing from non-ionizing radiation, individual atoms or molecules have ionization energies different from this (some, even lower). For example, the atoms of C, O and Li have ionization energies of 11.3 eV, 13.6 eV, and 5.4 eV, respectively. For the water molecule, the energy of ionization is ~12.6 eV.

3. INTERACTION OF NON-IONIZING EM RADIATION WITH MATTER

When interacting with matter, EM radiation can transfer energy to the atoms or molecules of that material. This process is called *absorption* of the EM radiation. As a result of absorption, the photon disappears, but transfers its energy to the absorbing atom or molecule.

As a result of absorption of non-ionizing EM radiation, the following processes can occur:

- increase of the material's temperature;
- excitation of atoms or molecules.

Not all radiation can be absorbed by any kind of atom or molecule! In order for the radiation to be absorbed, its energy must exactly match the difference between two energy levels in the atom or molecule. This is because the energy of an atom or molecule is *quantized*. Thus, any atom or molecule has certain characteristic energy levels.

The energy levels correspond to different states of motion of the entire atom or molecule (Figure 8.6) or to the energy of its electrons. These are, for a molecule[2]:

2 For an atom, only translational and electronic energy levels exist.

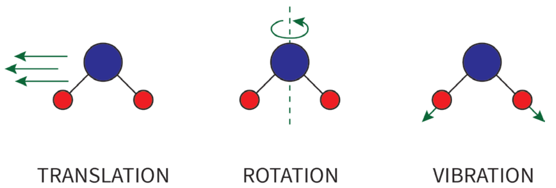


Figure 8.6. Different movement types in the water molecule. Left: translation = movement of the entire molecule in one direction of space. Middle: rotation around a certain axis. Right: vibration of the O–H bonds. Depicted is only one of several possible modes of vibration: a symmetric stretching of the O–H bonds where the two O–H bonds simultaneously elongate (and then simultaneously shorten).

- *translational levels* – correspond to the movement of the molecule in space (kinetic movement). Note that there are so many translational energy levels and the difference in energy between them is so small, that translational energy can generally be considered unquantized;
- *rotational levels* – correspond to the rotation of a molecule around an axis;
- *vibrational levels* – correspond to the movement of atoms in a chemical bond (stretching, bending of the bond);
- *electronic levels* correspond to the energy of the electrons inside the molecule.

We can thus, write that the total energy of a molecule can be calculated as:

$$E_{total} = E_{translation} + E_{rotation} + E_{vibration} + E_{electronic} \quad (8.7)$$

where:

$$E_{electronic} > E_{vibration} > E_{rotation} > E_{translation} \quad (8.8)$$

We can more easily visualize the differences between these energy levels by plotting them in an energy diagram where the energy increases from bottom to top, called a Jablonski diagram (Figure 8.7).

Under most conditions, atoms and molecules stay in a state of minimum energy called the *ground state* (noted as S_0 in Figure 8.7). When EM radiation is absorbed, the molecule goes into a state of higher energy. Depending on the energy of EM radiation absorbed, this can cause different types of transitions inside the molecule:

- microwave radiation causes transitions between rotational levels;
- IR radiation causes transitions between vibrational and rotational levels.
- when energy in the UV and visible (UV-VIS) domains is absorbed, an electron in the molecule can move up to a higher electronic level. We say that the molecule has become *excited*. We noted the first excited state of a molecule as S_1 in Figure 8.7, but higher energy excited states can also exist. Thus, **UV-VIS radiation excites transitions for all types of levels: electronic, vibrational and rotational**. Note that we ignored the transitional levels in this discussion, as they are so close together that transitions between them can be caused by radiation of very small energies.

The general process for a molecule in the ground state M that absorbs EM radiation in the UV-VIS domains, going into an excited state M* can be written as:

$$M + h\nu \rightarrow M^* \quad (8.9)$$

The absorption of EM radiation by atoms or molecules is studied by *spectroscopy*. The detailed representation of the degree of absorption of EM radiation as a function of its wavelength for a particular molecule is called the *absorption spectrum* of the respective molecule. An example of an absorption spectrum is shown later in this chapter (Figure 8.9). You will study the absorption of radiation in the UV and visible domains in more detail in the practical activities.

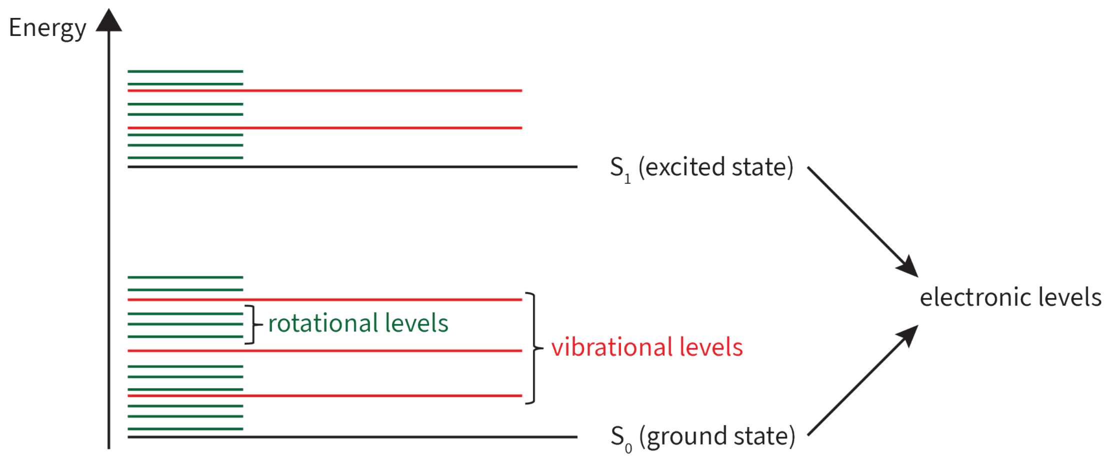


Figure 8.7. Jablonski diagram. The electronic, vibrational and rotational energy levels are shown. Translational levels are not shown; they would be a practically continuous spectrum between the rotational levels. The different lengths of the lines for the different types of levels were chosen for better visualization and have no physical meaning.

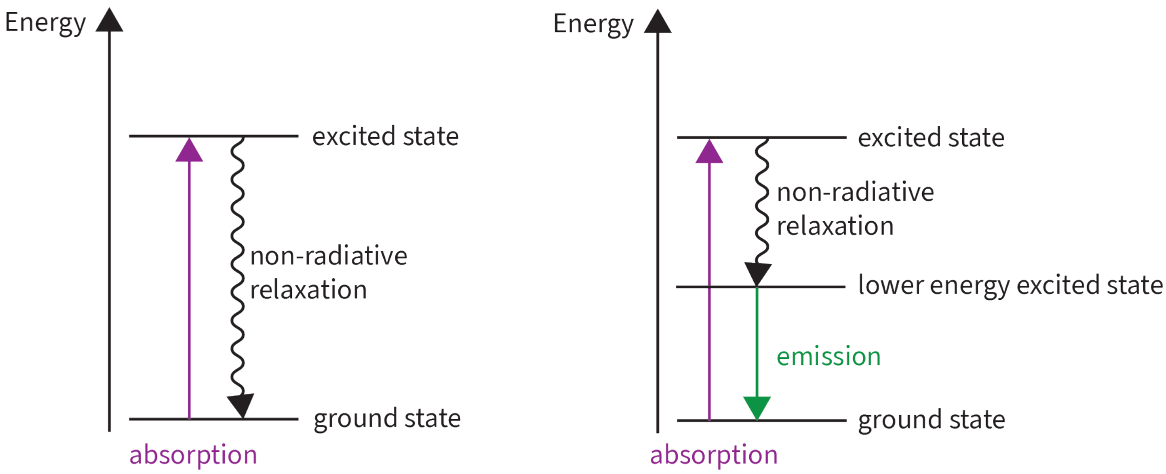


Figure 8.8. Simplified Jablonski diagram. Absorption of UV-VIS radiation by a molecule causes it to go from the ground state to an excited state. In most molecules (left panel), this process is followed by return to the ground state through non-radiative relaxation. For a small number of molecules (right panel), the return to ground state can also be accompanied by emission of a photon at a higher wavelength than the absorbed photon.

In the excited state, the molecule is not stable and it will try to return as soon as possible towards the ground state. This can be done through different types of processes that can generally be classified as:

- non-radiative relaxation – this returns the molecule to the ground state, for example, by transferring its energy to other molecules by collision. Overall, this results in the increase in the kinetic energy of the molecules, and thus an increase of the temperature;
- *emission* – this can happen only in a small number of substances. After absorption, the molecule can lose part of its energy through emitting another photon. **The emitted photon will be of lower energy than the absorbed photon,** thus, $\lambda_{emission} > \lambda_{absorption}$, as usually part of the energy absorbed is also lost by non-radiative relaxations. Emission is classified as:
  - *fluorescence* (rapid emission, on sub-nanosecond timescales);
  - *phosphorescence* (slow emission, on timescales of microseconds up to hours).

A simplified diagram of the absorption process and the return to the ground state is shown in Figure 8.8.

In the excited state, molecules can undergo special reactions that would normally not be possible in the ground state. These are called *photochemical reactions*. A few examples are listed below:

- the excited molecule reacts with a different molecule (A), a reaction which would not be possible in the ground state:

$$M^* + A \rightarrow B \qquad (8.10)$$

- the excited molecule forms a dimer:

$$M^* + M \rightarrow MM \qquad (8.11)$$

- the excited molecule transfers its energy to another molecule (A), that will then become reactive (*photosensitization*):

$$M^* + A \rightarrow M + A^* \qquad (8.12)$$

Photochemical reactions obey two laws:

- the **Grotthuss–Draper law**: in order for the photochemical reaction to take place, the molecule must absorb radiation;
- the **Stark–Einstein law**: for one photon that is absorbed, only a single molecule can react (the energy of one absorbed photon cannot be shared by different molecules).

4. BIOLOGICAL EFFECTS OF VISIBLE LIGHT

In general, all molecules can absorb UV radiation. However, only a limited number of molecules can absorb visible radiation. Chemical groups in molecules that absorb light in the visible domain are called *chromophores*.

Plants and other autotrophic organisms can use visible light emitted by the sun directly as an energy source in order to synthesize complex molecules (sugars), in a process called *photosynthesis*. In photosynthesis, sunlight is absorbed by a pigment (organic molecule) called chlorophyll, present in specialized organelles called chloroplasts. The general process of photosynthesis in which one molecule of glucose is synthesized can be written as:

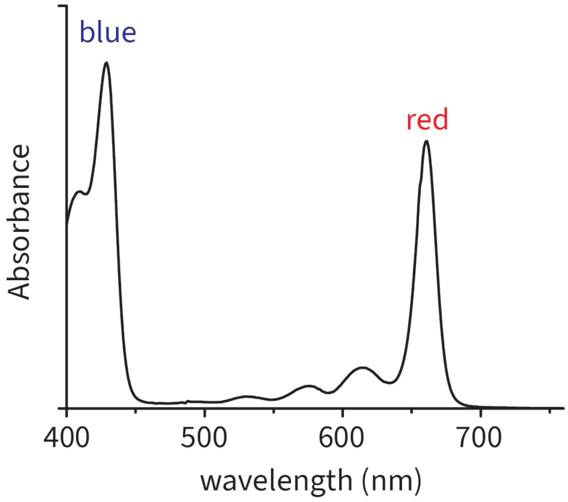


Figure 8.9. Absorption spectrum of chlorophyll a (the main form of chlorophyll in plants) in the visible range (400 – 760 nm).[3] Maxima of absorption are at ~430 nm (blue) and ~660 nm (red). No absorption of light occurs in the green part of the spectrum (~500 – 570 nm).

$$6CO_2 + 6H_2O \xrightarrow{h\nu} C_6H_{12}O_6 + 6O_2 \qquad (8.13)$$

Did you ever ask yourself why plants are green? Light emitted by the sun and other common light sources is *white light*: it contains all the possible wavelengths of the visible spectrum (all the colors of the rainbow: red, orange, yellow, green, blue, indigo, violet). If a material absorbs light in the visible spectrum, we will see it as having a particular color, which corresponds to the part of the visible spectrum that the respective material does not absorb (it reflects). For chlorophyll, the absorption spectrum (Figure 8.9) has maxima in the blue and red regions of the visible spectrum, while green light is not absorbed. Therefore, green light will be reflected and reach our eyes.

Once the green photons reflected by chlorophyll reach our eyes, they can be absorbed by specialized pigments in the photoreceptor cells in our retina; this is the first step in the process of vision (described in a separate chapter of this book).

Returning to color, we now have a good definition of what a material being colored tells us about the absorption spectrum of that material: the material will appear to have the color(s) that it does not absorb from the visible spectrum. As extreme cases, a black material absorbs at all wavelengths in the visible spectrum and a white material reflects all the light in the visible spectrum.

Let us now consider the following question – why are medicine bottles dark brown or opaque in color? We should now be able to answer this: the dark color of the bottle material will absorb visible radiation (and normally UV radiation as well, though you cannot tell that from the color), thus preventing light from reaching the molecules of the drug and causing unwanted photochemical reactions and degradation of the medicine. If we correlate this with what we learned from thermodynamics, we can easily understand why the general recommendation is to keep medicine away from light (we answered this already) and in a cool, dry place: an increase in temperature or humidity will make chemical reactions, and thus degradation of the medicine more likely. This is certainly undesired, as correct treatment requires administration of a precise dose of a particular drug. In addition, it is quite common for the degradation products of drugs to be toxic.

5. BIOLOGICAL EFFECTS OF UV RADIATION

UV radiation is the domain of EM radiation with wavelengths of (10 nm to 400 nm). As we saw before, radiations in the UV domain are either ionizing ($\lambda \leq 124$ nm) or non-ionizing ($\lambda > 124$ nm). UV radiation below ~200 nm is not of medical interest; it is rapidly absorbed by air molecules and can only propagate over long distances in a vacuum. It is, therefore, denoted as *vacuum UV*. Let us then focus on the biological effects of UV radiation with wavelengths between 200 and 400 nm. In this range, we can further subdivide UV radiation depending on its wavelength[4] as:

- UVA (315 nm – 400 nm);
- UVB (280 nm – 315 nm);
- UVC (200 nm – 280 nm).

In the range 200 nm to 400 nm the energies of UV photons vary between ~6 eV and ~3 eV. These are, thus, non-ionizing. However, exposure to UV radiation can have various deleterious effects on biological matter, as described below.

Exposure to UV radiation occurs naturally, from sunlight. Solar radiation directed at the Earth contains UVA, UVB and UVC radiation; however, the atmosphere (including the ozone layer in the stratosphere) absorbs all UVC radiation and most of the UVB radiation. Thus, UV radiation that reaches us from the sun contains mostly (90 – 95%) UVA and a small portion (5 – 10%) UVB. UV radiation is not highly penetrating – it is, thus, our skin and our eyes which can be exposed to

3 Figure was drawn using data provided by PhotochemCAD: Taniguchi, M. and Lindsey, J.S. (2018), Database of Absorption and Fluorescence Spectra of >300 Common Compounds for use in PhotochemCAD. *Photochem Photobiol*, 94: 290-327. *https://doi.org/10.1111/php.12860*

4 The exact limits of wavelength for UVA, UVB and UVC can vary slightly depending on the source.

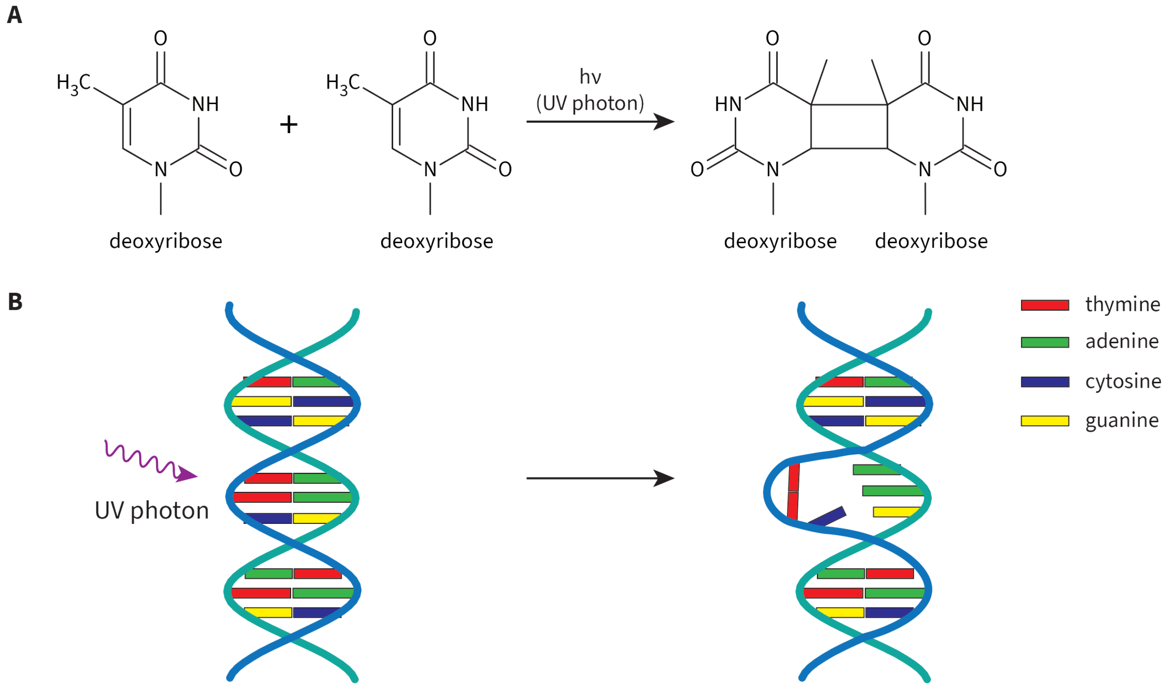


Figure 8.10. UV radiation can cause the dimerization of thymine. A, thymine dimerization reaction. B, effect of thymine dimerization on the DNA molecule.

UV radiation and suffer the consequences of this exposure.

5.1. Dimerization of pyrimidines

One of the main sources of damage to DNA following exposure to UV radiation is a photochemical dimerization of pyrimidine nucleotide bases in DNA, thymine and cytosine. Thus, in the presence of UVB or UVC radiation, pyrimidine molecules that are adjacent to each other on a DNA strand can react, forming a pyrimidine dimer (Figure 8.10). As a result of pyrimidine dimerization, the DNA molecule is damaged, altering its transcription and potentially preventing replication. Furthermore, a defective replication of the affected DNA can lead to mutations, resulting in the development of skin cancer.

The human body has a means of repairing the damaged DNA regions by removing them altogether, in a process called nucleotide excision repair. This is, of course, not foolproof and excessive exposure to UV radiation will still damage the cells. A genetic disorder called *xeroderma pigmentosum* is caused by mutations that partially or completely damage the nucleotide excision repair enzymes. Patients suffering from this condition are extremely sensitive to UV radiation, suffering from severe sunburn upon exposure to the sun, a high probability of skin cancer and cataracts. There is no known cure, and patients must strictly avoid exposure to sunlight.

Table 8.1. Examples of reactive oxygen species.

Name	Chemical formula
Superoxide ion	O_2^-
Hydroperoxyl	HOO·
Hydroxyl radical	HO·
Ozone	O_3
Hydrogen peroxide	H_2O_2

5.2. Formation of reactive oxygen species (ROS)

Reactive oxygen species (ROS) are chemical species containing oxygen that have a very high chemical reactivity. Some of the most common ROS are hydrogen peroxide (H_2O_2), superoxide ions (O_2^-) and hydroxyl radicals (HO·). A few examples of ROS are given in Table 8.1. As you can see, some ROS are free radicals (molecules with one or more unpaired electrons), but can also be non-radicals. Note that **water and the species resulting from the auto-dissociation of water (HO^- and H_3O^+) are not ROS.**

Both UVA and UVB radiation have the capability of forming ROS. While the damage of UVB is limited to the epidermis, as it is less penetrating, UVA also passes into the dermis, leading to deeper damage. Increase of ROS following exposure to UV radiation will cause oxidative stress. This can lead to cellular damage through lipid peroxidation and fragmentation of DNA.

As a combined result of the effects of UV radiation (thymine dimerization, ROS formation), over time, *photoaging* appears – a degradation of elastic and collagen fibers in the dermis.

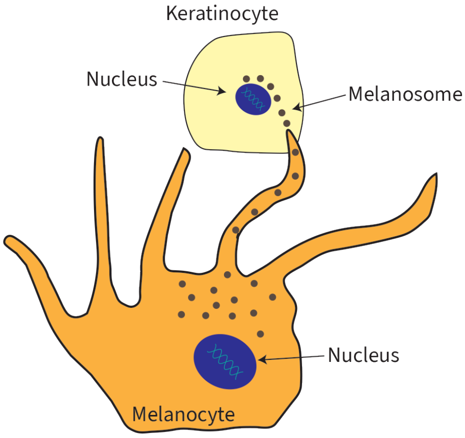


Figure 8.11. A melanocyte releases melanosomes in a keratinocyte in order to provide protection from UV radiation.

5.3. Skin pigmentation (tanning)

Since UV radiation is damaging, the skin has a natural defense mechanism: sun tanning.

Keratinocytes are the main type of cell found in the epidermis. Melanocytes (Figure 8.11) are found in the basal layer of the epidermis and play a role in protecting keratinocytes (and, thus, the skin), from the effects of UV radiation, through the production of a dark pigment called *melanin*.

Melanin is the pigment responsible for the different skin colors and is stored in *melanosomes*. It is synthesized from the amino acid tyrosine. While melanocytes continuously synthetize melanin, this process is up-regulated following exposure to UV radiation (UVB and UVA). Melanin is transferred to keratinocytes from the melanocytes and serves a protective role, by absorbing incoming UV radiation and converting it into heat through non-radiative relaxation.

5.4. Sunburn and actinic keratosis

The protection offered by melanin to the skin is not complete. Overexposure to UVB or UVC radiation causes an acute response in the skin known as *sunburn*. Sunburn is characterized by erythema (redness of the skin), edema (swelling) and increased tissue temperature. Sunburn is an inflammatory response caused by skin cells being damaged by UV radiation and starting to die via apoptosis (controlled death), releasing inflammatory cytokines. Severe pain also accompanies sunburn. In general, symptoms recede after a few days, but sunburn is associated with an increased risk of skin cancers.

Following long-term exposure to UV radiation, keratinocytes can start growing in an abnormal fashion, leading to scaly patches of skin called *actinic keratosis*. These are pre-cancerous lesions that may, in time, progress to squamous cell carcinoma.

5.5. Photokeratitis

Photokeratitis refers to the damage produced by UV radiation to the cornea. It can be likened to a sunburn of the eye. Intense exposure of the eye to UV radiation causes damage to the epithelium of the cornea. Most commonly, this occurs in welders working without protective equipment or in high altitude settings (skiing, mountain climbing in snow) where UV radiation is reflected by the snow. The condition also has common names depending on the causes: “snow blindness” or “welder’s flash”. Symptoms of photokeratitis include pain, excessive tear production, photoaversion and decreased vision. Normally, symptoms subside within 1 – 3 days as the corneal epithelium regenerates.

5.6. Germicidal effect of UVC radiation

Even though UVC radiation does not reach the surface of the Earth through sunlight, it can be produced via artificial means (UVC lamps). Most commonly, UVC lamps employ a wavelength of 254 nm, as the absorption spectrum of DNA has a maximum at ~260 nm. UVC lamps are commonly used in food production, research labs, hospitals, etc. for their germicidal activity, as UVC has a damaging effect to genetic material (more powerful than UVB). The germicidal effect of UVC extends to bacteria, fungi, viruses (such as SARS-CoV-2), etc.

Care should be taken when employing UVC lamps that humans are not exposed to UVC radiation. Accidental exposure can lead to erythema and photokeratitis.

5.7. Protection from UV radiation

We’ll explicitly point out here what should be obvious – exposure to UV radiation is severely damaging to the skin and should be avoided. UV radiation is the main cause of melanoma and non-melanoma (basal cell carcinoma and squamous cell carcinoma) skin cancer. Exposure to UV radiation also contributes to the formation of *cataracts*: clouding of the crystalline lens of the eye.

The main source of exposure to UV radiation is natural: outdoor exposure to sunlight. Note that exposure to UV radiation occurs also on cloudy days, as clouds do not completely block UV radiation. The amount of UV radiation exposure from sunlight on a particular day is evaluated by the UV index, which you can consult when checking the weather. Indoor, window glass blocks the majority

of UVB radiation, but usually allows UVA to pass. Artificial sources of UV radiation are also a potential risk. For instance, fluorescent light bulbs emit a small amount of UVA and UVB radiation.

A more serious risk is represented by the use of tanning beds that employ UVA radiation. UV exposure has been shown to promote synthesis of the endogenous opioid β-endorphin in the skin. This has made some physicians push for excessive tanning to be classified as an addiction. Whether addictive or not, though, tanning is certainly not healthy for the skin. Remember that tanning is our body's defense response against UV radiation, not something that is to be desired, as even the increased melanin production is not enough to fully protect skin from the effects of UV radiation.

Finally, another common artificial source of UV radiation (UVA) is represented by nail polish dryers that are used to fully cure UV-sensitive nail gels. Recently, it has been shown that exposure to UVA radiation from nail polish dryers kills human and mouse cells *in vitro*. While this does not show direct correlation between nail polish dryers and skin cancers of the hand, and requires further study, it might be prudent to avoid excessive use of such devices in the meantime.

If exposure to UV radiation from sunlight cannot be avoided, sunscreen of adequate protection factor should be used. The sun protection factor (SPF) is used to characterize sunscreens. The SPF is calculated as the ratio between the least amount of UV radiation required to produce a minimal erythema on skin protected by sunscreen and the amount of energy required to produce the same erythema on bare, unprotected skin:

$$SPF = \frac{MED\ (with\ sunscreen)}{MED\ (without\ sunscreen)} \quad (8.14)$$

where MED = minimal erythemal dose.

As erythema is mainly caused by UVB, the SPF only shows the amount of protection from UVB. The percentage of UVB radiation absorbed by the sunscreen can be calculated from the SPF as:

$$absorption\ (\%) = 100 - \frac{100}{SPF} \quad (8.15)$$

Therefore, a sunscreen with SPF15 absorbs ~93% of all UVB radiation, a sunscreen with SPF 30 absorbs ~97% of UVB, and a sunscreen with SPF50 absorbs 98% of UVB. Care should be taken that enough sunscreen is applied – the SPF is measured when 2 mg of sunscreen is applied over 1 cm^2 of skin, while most people apply much less. The sunscreen should also be reapplied periodically (every ~2 h).

It is currently recommended that sunscreens of at least SPF30 – SPF50 should be used, and to ensure that the used sunscreen also provides protection from UVA radiation, as UVA also has damaging effects on the skin. Remember that protection from UVA is not included in the SPF value, so you should look up explicitly on the label of the sunscreen if it also includes UVA protection (in the US, this is labelled as "broad spectrum" protection).

Should you care about anything other than SPF and UVA protection when choosing a sunscreen? Yes, the chemical composition is also important. Classical formulations of sunscreen contain organic molecules that are absorbed into the skin. When exposed to UV these degrade through photochemical reactions, forming chemical byproducts that may be irritating or even toxic. The most concerning ingredient is oxybenzone, which is a hormone disrupter.

Formulations of sunscreen that contain nanoparticles of titanium dioxide (TiO_2) and zinc oxide (ZnO) are not absorbed, but form a protective layer on top of the skin. As a result, at present, only TiO_2 and ZnO are designated as GRASE (generally recognized as safe and effective) by the US Food and Drug Administration.

5.8. Synthesis of vitamin D_3 in the skin

We have, so far, listed, only negative, damaging effects of UV radiation on the human body. However, UV radiation also has a positive effect. Exposure of the skin to UVB promotes the synthesis of *cholecalciferol* (vitamin D_3) in the skin from 7-dehydrocholesterol. Cholecalciferol is converted in the liver to calcifediol, which, in turn, is converted in the kidneys to calcitriol (vitamin D). The exact reactions will be studied in Biochemistry. Calcitriol promotes calcium absorption in the intestine; deficiency of vitamin D leads to impaired bone development and maintenance. If children are deficient in vitamin D, it will lead to a condition called *rickets*, where bones are softer and weaker than normal.

Maintaining the amounts of vitamin D_3 needed by the body requires either regular exposure to sunlight or taking dietary supplements containing vitamin D. This exposure can be as little as a few minutes per day, but can vary depending on the time of day, season of the year, or color of the person's skin (darker shades of skin require more exposure to UVB to generate the same amount of vitamin D_3 as the higher amount of melanin blocks more UVB radiation than in individuals with lighter skin). Thus, in summer, the time of UV exposure needed for synthesis of vitamin D might be as low as 5 – 10 minutes, whereas, in winter, when the UV levels are lower and a much smaller fraction of our skin is uncovered, it might increase to several hours. As such, in winter,

taking of vitamin D supplements is advised to prevent vitamin D deficiency.

5.9. Phototherapy

UV radiation can be employed in the treatment of some skin conditions such as eczema, psoriasis or vitiligo. This is described in the chapter Physical factors in therapy.

REFERENCES

Atkins, P. W., De Paula, J., & Keeler, J. (2017). *Atkins' Physical Chemistry*. London: Oxford University Press.

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Bennett, S. L., & Khachemoune, A. (2022). Dispelling myths about sunscreen. *J Dermatolog Treat, 33*(2), 666-670. doi:10.1080/09546634.2020.1789047

Bishop, T. (2007). UV-Induced Erythema. In R. F. Schmidt & W. D. Willis (Eds.), *Encyclopedia of Pain* (pp. 2595-2597). Berlin, Heidelberg: Springer Berlin Heidelberg.

Brenner, M., & Hearing, V. J. (2008). The protective role of melanin against UV damage in human skin. *Photochem Photobiol, 84*(3), 539-549. doi:10.1111/j.1751-1097.2007.00226.x

Cromwell, B. (2010). *Light and Matter*. Retrieved from http://www.lightandmatter.com/lm/

Dale Wilson, B., Moon, S., & Armstrong, F. (2012). Comprehensive review of ultraviolet radiation and the current status on sunscreens. *J Clin Aesthet Dermatol, 5*(9), 18-23. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23050030

de Jager, T. L., Cockrell, A. E., & Du Plessis, S. S. (2017). Ultraviolet Light Induced Generation of Reactive Oxygen Species. *Adv Exp Med Biol, 996*, 15-23. doi:10.1007/978-3-319-56017-5_2

D'Orazio, J., Jarrett, S., Amaro-Ortiz, A., & Scott, T. (2013). UV radiation and the skin. *Int J Mol Sci, 14*(6), 12222-12248. doi:10.3390/ijms140612222

Fell, Gillian L., Robinson, Kathleen C., Mao, J., Woolf, Clifford J., & Fisher, David E. (2014). Skin β-Endorphin Mediates Addiction to UV Light. *Cell, 157*(7), 1527-1534. doi:https://doi.org/10.1016/j.cell.2014.04.032

Flowers, P., Theopold, K., Langley, R., & Robinson, W. R. (2019). *Chemistry 2e*. Retrieved from https://openstax.org/books/chemistry-2e/pages/1-introduction

Franklin, K., Muir, P., Scott, T., & Yates, P. (2019). *Introduction to Biological Physics for the Health and Life Sciences*: Wiley.

Izadi, M., Jonaidi-Jafari, N., Pourazizi, M., Alemzadeh-Ansari, M. H., & Hoseinpourfard, M. J. (2018). Photokeratitis induced by ultraviolet radiation in travelers: A major health problem. *J Postgrad Med, 64*(1), 40-46. doi:10.4103/jpgm.JPGM_52_17

Leung, A. K., Barankin, B., Lam, J. M., Leong, K. F., & Hon, K. L. (2022). Xeroderma pigmentosum: an updated review. *Drugs Context, 11*. doi:10.7573/dic.2022-2-5

Linetsky, M., Raghavan, C. T., Johar, K., Fan, X., Monnier, V. M., Vasavada, A. R., & Nagaraj, R. H. (2014). UVA light-excited kynurenines oxidize ascorbate and modify lens proteins through the formation of advanced glycation end products: implications for human lens aging and cataract formation. *J Biol Chem, 289*(24), 17111-17123. doi:10.1074/jbc.M114.554410

Narbutt, J., Philipsen, P. A., Harrison, G. I., Morgan, K. A., Lawrence, K. P., Baczynska, K. A., . . . Young, A. R. (2019). Sunscreen applied at >/= 2 mg cm(-2) during a sunny holiday prevents erythema, a biomarker of ultraviolet radiation-induced DNA damage and suppression of acquired immunity. *Br J Dermatol, 180*(3), 604-614. doi:10.1111/bjd.17277

Reardon, J. T., & Sancar, A. (2003). Recognition and repair of the cyclobutane thymine dimer, a major cause of skin cancers, by the human excision nuclease. *Genes Dev, 17*(20), 2539-2551. doi:10.1101/gad.1131003

Religi, A., Backes, C., Chatelan, A., Bulliard, J. L., Vuilleumier, L., Moccozet, L., . . . Vernez, D. (2019). Estimation of exposure durations for vitamin D production and sunburn risk in Switzerland. *J Expo Sci Environ Epidemiol, 29*(6), 742-752. doi:10.1038/s41370-019-0137-2

Sample, A., & He, Y. Y. (2018). Mechanisms and prevention of UV-induced melanoma. *Photodermatol Photoimmunol Photomed, 34*(1), 13-24. doi:10.1111/phpp.12329

Taniguchi, M., & Lindsey, J. S. (2018). Database of Absorption and Fluorescence Spectra of >300 Common Compounds for use in PhotochemCAD. *Photochemistry and Photobiology, 94*(2), 290-327. doi:https://doi.org/10.1111/php.12860

United States Environmental Protection Agency. Calculating the UV Index. Retrieved from https://www.epa.gov/sunsafety/calculating-uv-index-0

United States Environmental Protection Agency. UV Index Scale. Retrieved from https://www.epa.gov/sunsafety/uv-index-scale-0

Wang, X., Kinziabulatova, L., Bortoli, M., Manickoth, A., Barilla, M. A., Huang, H., .

. . . Lumb, J.-P. (2023). Indole-5,6-quinones display hallmark properties of eumelanin. *Nature Chemistry, 15*(6), 787-793. doi:10.1038/s41557-023-01175-4

Zhivagui, M., Hoda, A., Valenzuela, N., Yeh, Y.-Y., Dai, J., He, Y., . . . Alexandrov, L. B. (2023). DNA damage and somatic mutations in mammalian cells after irradiation with a nail polish dryer. *Nature Communications, 14*(1), 276. doi:10.1038/s41467-023-35876-8

CHAPTER 9

RADIOBIOLOGY

Prerequisite knowledge

- Structure of the atom
- Mass-energy equivalence
- Electromagnetic radiation

1. RADIOBIOLOGY. CLASSIFICATION OF IONIZING RADIATION

1.1. Ionizing radiation and radiobiology

As we have seen in the chapter on Photobiology, we call ionizing radiation any radiation that is capable of *ionizing* atoms or molecules by extracting electrons from their structure (Figure 8.5 of the previous chapter). **Radiation with energies above 10 eV is considered ionizing radiation**.

Radiobiology studies the interaction of ionizing radiation with biological systems.

1.2. Classification of ionizing radiation

Depending on their nature, we can classify ionizing radiation as:

- **electromagnetic (EM) radiation**: γ radiation, X-rays and UV radiation with λ ≤ 124 nm;
- **particle radiation** (composed of subatomic scale or larger particles) that can be further classified as:
  - charged particles: α radiation, β radiation, protons, heavy ions, etc.;
  - uncharged particles: neutrons.

2. SOURCES OF IONIZING RADIATION

2.1. Natural sources

You might be surprised to learn that we are, every day, exposed to ionizing radiation from natural sources. This is called the *natural background radiation*. This is not a cause for concern – the exact dose received annually from natural sources is low, but we'll give some numbers later, after we discuss the means of expressing doses.

Figure 9.1 shows the main sources of ionizing radiation that form the natural background, according to a report of the US National Council on Radiation Protection & Measurements (NCRP).

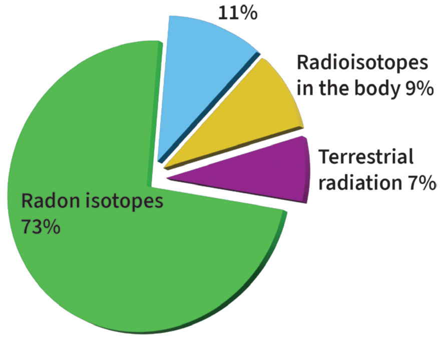


Figure 9.1. Exposure to natural sources of ionizing radiation[1]. Data show the average exposure of US citizens in 2006 to background ionizing radiation.

The main source of ionizing radiation from the natural background comes from two isotopes of the gas radon: ^{222}Rn (usually called just "radon") and, in a smaller proportion, ^{220}Rn (usually called "thoron"). These result from the radioactive decay of naturally occurring *radioactive isotopes* in the Earth's crust. Mainly, exposure to ionizing radiation from radon isotopes occurs indoors, as they are released through the radioactive decay of radioactive isotopes in the materials that the building is made from (concrete, bricks, etc.).

A small part of the natural background comes from the sun in the form of *cosmic rays*: particle radiation that is mostly blocked by the atmosphere. Exposure to cosmic rays increases the higher one goes in altitude, for example during an airplane flight.

An even smaller contribution comes from radioisotopes that are naturally present in our body. The main source is ^{40}K, a naturally occurring radioactive isotope of potassium with an abundance of ~0.01% of all potassium found in nature. Thus, ^{40}K is always present in our bodies in this small proportion, contributing to a constant internal exposure to ionizing radiation.

Finally, terrestrial radiation refers to all the naturally occurring radioisotopes present in the

1 Data source: National Council on Radiation Protection & Measurements. *(2009). NCRP Report No. 160, Ionizing Radiation Exposure of the Population of the United States*.

crust of the Earth. These decay, releasing ionizing radiation and also usually producing another radioactive isotope that, in turn, will also decay.

2.2. Artificial sources

The main, and usually only source of exposure to ionizing radiation of artificial origin for an average person comes from medical procedures. These encompass either imaging procedures (radiography, CT scans, PET scans, etc.) or radiotherapy. In 2006, the aforementioned report of the NCRP indicated that average exposure of US citizens to ionizing radiation from medical sources was roughly equal to the exposure from natural sources. In 2016, also according to the NCRP, exposure to ionizing radiation from medical sources decreased by ~26% compared to 2006, which is attributable to the development of newer radiography and CT machines that reduce the dose the patients are exposed to.

Artificial exposure can also occur at the workplace for individuals such as: miners, scientists working with radioactive isotopes, power plant workers, soldiers, etc.

A common confusion that propagates in the media is that phones expose us to “radiation” that the public is lead to believe to be comparable to radiation emitted by radioactive sources. This is incorrect. The **radiation used for communication in cell phones and wireless networks is NOT ionizing radiation**, but is in the microwave – radio wave range, and thus non-ionizing.

3. RADIOACTIVITY

3.1. Isotopes. Radioactive isotopes

We used the term “radioactive isotope” throughout the previous section. Let us briefly explain what that is. The atomic nucleus is composed from protons and neutrons. Protons are electrically charged, each of them carrying one elementary charge. As such, a nucleus that would contain only protons would be unstable because of the powerful electrical repulsion that would occur. The only stable nucleus composed of only protons is that of the hydrogen atom, which has only 1 proton. We can, thus, conclude that neutrons stabilize the nucleus and allow it to contain more protons.

For a given chemical element, the name of the element corresponds to the *atomic number*, Z, which is the number of protons. For example, for Z = 6, the element’s name is carbon. However, several chemical species exist that have 6 protons in the nucleus, but different numbers of neutrons (6, 7, 8 and so on.). These are all species of carbon that we call *isotopes* of carbon. We denote the different number of neutrons by calculating the number of atomic mass (A):

$$A = no.\ of\ protons\ (Z) + no.\ of\ neutrons \qquad (9.1)$$

To denote the nuclear structure of an isotope of a particular element whose chemical symbol is X, we can write it as ${}^{A}_{Z}X$, or simply as ${}^{A}X$. For example, carbon has three isotopes that are found in nature: ${}^{12}_{6}C$, ${}^{13}_{6}C$, and ${}^{14}_{6}C$, which we pronounce when reading as “carbon-twelve”, “carbon-thirteen”, “carbon-fourteen” (Figure 9.2). Note that writing the number of protons in the lower left is optional, as it confers the same information as the symbol of the chemical element.

For a netural atom, the number of protons in the nucleus is equal to the number of electrons in the electron shell. Therefore, all isotopes of the same element will have the same number of electrons. Consequently, as chemical reactivity is given by the number and distribution of electrons, isotopes of the same element have the same chemical properties. So, all the three isotopes of carbon in Figure 9.2 will give the same chemical reactions and are all found in natural carbon sources, in different proportions.

There is, however, one large difference between the stability of the nuclei of those three isotopes: the ${}^{14}C$ nucleus is unstable (simply speaking, it has

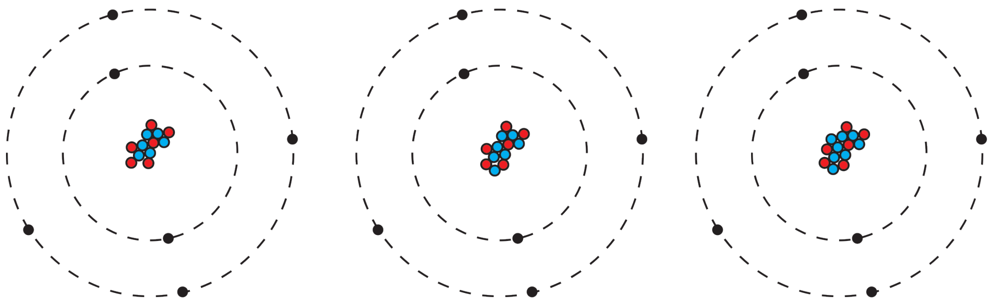


Figure 9.2. Natural isotopes of carbon and their relative abundance. Red circle = proton; blue circle = neutron; black circle = electron. Thus, ${}^{12}C$ has 6 protons and 6 neutrons, ${}^{13}C$ has 6 protons and 7 neutrons, while ${}^{14}C$ has 6 protons and 8 neutrons.

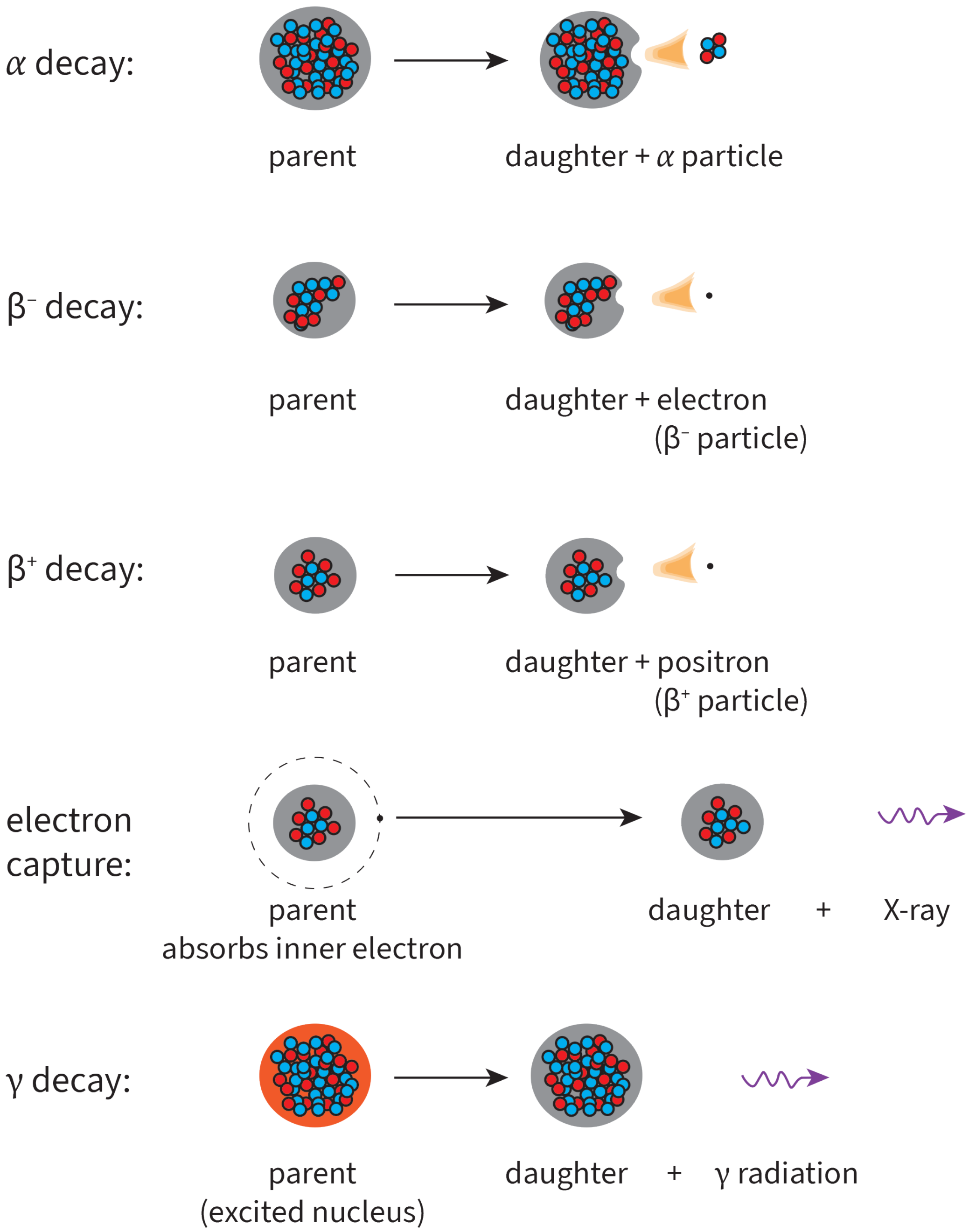


Figure 9.3. Types of radioactive decay. Only the ionizing radiation emitted is depicted. Description is in the text. Red circle = proton; blue circle = neutron; black circle = electron or positron.

too many neutrons relative to the number of protons), and there is a chance for it to spontaneously disintegrate, releasing ionizing radiation in the process. This disintegration is called **radioactive decay**. Isotopes with unstable nuclei are called **radioactive isotopes** (**radioisotopes**). Isotopes such as ^{12}C or ^{13}C whose nuclei do not disintegrate spontaneously are called **stable isotopes**.

Generally speaking, all chemical elements can have more than one isotope: some only a few while others have tens of isotopes. There are some elements which have only a single stable isotope (e.g. Be, F, Al, etc.), while others have no stable isotopes (generally, those with $Z > 82$).

^{14}C is an example of a *natural radioisotope*, it is produced naturally in the atmosphere from nitrogen through the action of cosmic rays. We can also produce *artificial radioisotopes* by bombarding nuclei with ionizing radiations (usually, neutrons). Artificial radioisotopes are commonly used in medical applications (imaging or radiotherapy).

3.2. Types of radioactive decay

The radioactive process can generally be described as follows: a radioactive nucleus (*radionuclide*) called the *parent nuclide* decays into a *daughter nuclide* and, in the process, loses part of its energy

that it releases in the form of ionizing radiation.

The most common types of radioactive decay are alpha (α), beta minus (β^-), beta plus (β^+), electron capture and gamma (γ) decay (Figure 9.3). We will discuss them in turn.

▸ ***α* decay** occurs in large radioactive nuclei and results in the emission of an α particle, consisting of 2 protons and 2 neutrons (a nucleus of He). The process can be written as:

$$^{A}_{Z}X \rightarrow {}^{A-4}_{Z-2}Y + {}^{4}_{2}He \tag{9.2}$$

where $^{A}_{Z}X$ is is the parent nuclide, $^{A-4}_{Z-2}Y$ is the daughter and $^{4}_{2}He$ is the α particle (ionizing radiation).

In the α decay process, the parent nucleus disappears. In its place, a smaller daughter nucleus remains, that has a mass lower than that of the parent by 4 atomic mass units. The α particle emitted also has an electric charge of +2, so it should be fully written as $^{4}_{2}He^{2+}$. However, the charge of the α particle is usually not depicted in decay reactions such as the one above, as these only concern themselves with the nucleus, and not the outer electron shell.

An example of α decay is the decay of ^{226}Ra, which results in the formation of ^{222}Rn:

$$^{226}_{88}Ra \rightarrow {}^{222}_{86}Rn + {}^{4}_{2}He \tag{9.3}$$

▸ **β⁻ decay** occurs in nuclei that have more neutrons than stable isotopes of the same element. In this decay process, a neutron turns into a proton and an electron: the proton remains in the nucleus of the daughter while the electron (β^- particle) is ejected as ionizing radiation. Additionally, another small particle called an *antineutrino* is emitted. The antineutrino has no charge, almost zero mass and interacts very weakly with matter. It is, therefore, not ionizing radiation:

$$n \rightarrow p^+ + e^- + \tilde{\nu} \tag{9.4}$$

where n is a neutron, p^+ is a proton, e^- is an electron (β^- particle) and $\tilde{\nu}$ is an antineutrino.

The overall process can be summarized as:

$$^{A}_{Z}X \rightarrow {}_{Z+1}^{A}Y + e^- + \tilde{\nu} \tag{9.5}$$

where $^{A}_{Z}X$ is the parent nuclide and $_{Z+1}^{A}Y$ is the daughter nuclide.

Overall, the daughter retains the atomic mass number of the parent, but increases its atomic number by 1. An example of a β^- decay is the radioactive decay of ^{14}C to ^{14}N:

$$^{14}_{6}C \rightarrow {}^{14}_{7}N + e^- + \tilde{\nu} \tag{9.6}$$

▸ **β⁺ decay** occurs in radionuclides that have too many protons compared to a stable isotope of the same atomic mass. In this decay process, a proton turns into a neutron and a *positron* (β^+ particle). The newly formed neutron remains in the daughter's nucleus (thus, the atomic mass stays the same as for the parent), while the positron is ejected as ionizing radiation. Additionally, a *neutrino* is formed. Just as its antiparticle (the antineutrino), the neutrino is not ionizing radiation.

$$p^+ \rightarrow n + e^+ + \nu \tag{9.7}$$

where p^+ is a proton, n is a neutron, e^+ is a positron (β^+ particle) and ν is a neutrino.

The process can be summarized as:

$$^{A}_{Z}X \rightarrow {}_{Z-1}^{A}Y + e^+ + \nu \tag{9.8}$$

where $^{A}_{Z}X$ is the parent nuclide and $_{Z-1}^{A}Y$ is the daughter nuclide.

Overall, the daughter retains the atomic mass number of the parent, but lowers its atomic number by 1. An example of a β^- decay is the radioactive decay of ^{18}F to ^{18}O:

$$^{18}_{9}F \rightarrow {}^{18}_{8}O + e^+ + \nu \tag{9.9}$$

The positron is the *antiparticle* of the electron: it has the same mass, but opposite electric charge and magnetic moment. When a positron meets an electron, an annihilation event occurs, where both particles disappear and their energy is converted into two γ photons. This is exploited in positron emission tomography (PET) imaging, as we will see in a future chapter.

It might be confusing how a neutron can turn into a proton (or the reverse). This is true because both of these are not elementary particles: they are made up of even smaller particles called *quarks*. Also, it might appear that mass is not conserved in the processes of β^- and β^+ decay. Remember that mass can be converted to energy (and the reverse). The total energy is always conserved in these decay processes.

▸ An additional, more rare type of radioactive decay can occur in nuclei that undergo β^+ emission. This is called **electron capture** and involves the capture of an electron from the inner electron shell by a proton, forming a neutron and a neutrino:

$$p^+ + e^- \rightarrow n + \nu \tag{9.10}$$

Overall, the process can be summarized as:

$$^{A}_{Z}X + e^- \rightarrow {}_{Z-1}^{A}Y + \nu \tag{9.11}$$

where $^{A}_{Z}X$ is the parent nuclide and $_{Z-1}^{A}Y$ is the

daughter nuclide.

We can see that electron capture results in the same daughter nuclide as that of β^+ decay, but without emission of a positron. Instead, the ionizing radiation emitted is usually an X-ray that results from a higher energy electron moving down in energy to fill up the vacancy left in the inner shell by the electron capture process. An example of electron capture occurs in ^{40}K. While, most of the time (~89%), ^{40}K decays via β^- emission to ^{40}Ca, about ~11% of the time, it decays via electron capture to ^{40}Ar:

$$^{40}_{19}K + e^- \rightarrow {}^{40}_{18}Ar + \nu \qquad (9.12)$$

▸ Finally, **γ decay** is a radioactive decay process in which a nucleus in an excited state goes to a lower energy state while emitting a γ photon. γ decay usually occurs immediately after the other decay processes (α, β^-, β^+, electron capture), as the daughter nucleus resulting from these processes is usually in an excited state. γ decay can be summarized as:

$$^{A}_{Z}X^* \rightarrow {}^{A}_{Z}X + \gamma \qquad (9.13)$$

where $^{A}_{Z}X^*$ is a nucleus in an excited state, $^{A}_{Z}X$ is a nucleus in a lower energy state and γ is a gamma photon.

The types of ionizing radiation produced through the presented processes and their characteristics are summarized in Table 9.1.

3.3. The law of radioactive decay. Half-life

Each individual radioactive nucleus has a certain probability of undergoing spontaneous decay at any point in time. This does not depend on the chemical state of the nuclide (charge, what bonds it forms etc.), nor does it depend on temperature. While it is impossible to predict when, exactly, a certain individual nucleus will decay, if a large enough number of radionuclides are present, the population can be analyzed statistically.

Radioactive decay follows what we call in chemistry a first order kinetics (a reaction that depends only on the concentration of one reactant). If we consider that we start at a particular point in time ($t = 0$) with a number N_0 of radionuclides of a particular isotope, we can calculate the remaining number of radionuclides, N(t) after a given time *t* according to an exponential decay equation called the *radioactive decay law*:

$$N(t) = N_0 e^{-\lambda t} \qquad (9.14)$$

where N_0 is the initial number of radioactive nuclei, N(t) is the number after the time t, e is the base of the natural logarithm (Euler's number, e = 2.718...), t is the time elapsed and λ is the radioactive decay constant.

For each radioisotope, we can measure a *radioactive decay constant*, noted as λ, which shows how likely it is for that particular isotope to decay: **higher values of λ mean a more unstable radioisotope**. is measured in units of time to the power of –1 (s^{-1}, min^{-1}, $years^{-1}$, etc.).

It is usually more convenient to express the stability of a given radioisotope by a related constant, called the *half-life* ($t_{1/2}$). For a radioisotope, the half-life represents the time that it will take for a given number of nuclei N_0 to decrease to half (to $N_0/2$). It can be calculated that this number is independent of the exact value of N_0 and only depends on the decay constant, according to:

$$t_{1/2} = \frac{\ln 2}{\lambda} = \frac{0.693}{\lambda} \qquad (9.15)$$

Thus, half-life is inversely proportional to the decay constant. It is measured in units of time (s, min, years, etc.). **The higher the half-life, the more stable a radioisotope is**. Half-life can vary wildly between different radioisotopes: some have half-lives of billions of years (^{40}K, $t_{1/2} = 1.3 \cdot 10^9$ years), while others have half-lives of fractions of

Table 9.1. Types of ionizing radiation produced through radioactive decay.

Type of radiation	Physical nature	Mass (amu)*	Charge
α particles	nuclei of He (^4_2He)	4	+2
β^- particles	electrons	0.00055	–1
β^+ particles	positrons	0.00055	+1
X-rays, γ rays	EM radiation (photons)	0	0

* 1 amu (atomic mass unit) = $1.67 \cdot 10^{-27}$ kg (1⁄12 of the mass of an atom of ^{12}C).

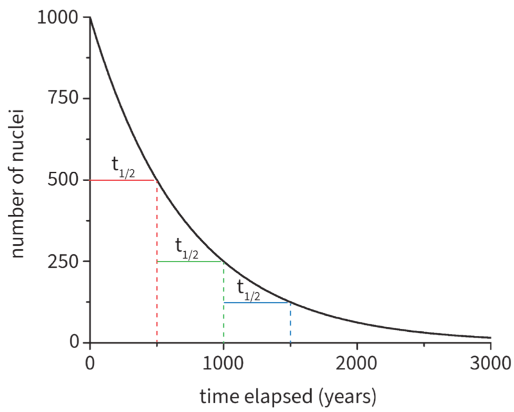


Figure 9.4. Radioactive decay and half-life. The graph shows the decay of a radioisotope with a half-life of 500 years, starting from N_0 = 1000 nuclei. Note that the half-life is the same, no matter whether we consider the time to go from 1000 to 500 nuclei (red line), from 500 to 250 (green line), or from 250 to 125 (blue line).

a second (^{219}Th, $t_{1/2}$ = 1 μs).

As the decay constant and the half-life are related only through a constant, we can equally use either of them to characterize the stability of a radioisotope. The radioactive decay law is shown graphically in Figure 9.4.

3.4. Activity of a radioactive source

For any radioactive source we can define its *activity* as the number of nuclei that decay per unit of time:

$$A = -\frac{dN}{dt} = \lambda N \qquad (9.16)$$

where the minus sign shows that the number of parent nuclei decreases, N is the number of nuclei at a given point in time, and λ is the decay constant.

In the International System of Units (SI), activity is measured in disintegrations/second, which is equivalent to saying s^{-1} as "disintegrations" is really a dimensionless number. This unit carries the name **becquerel (Bq)**. An older unit (we will call this a *legacy unit*) still in use is the **curie (Ci)**, which is the activity of 1 g of ^{226}Ra: 1 Ci = $3.7 \cdot 10^{10}$ s^{-1} = 37 GBq.

3.5. Radioactive decay chains

Heavy radioactive isotopes have been present in the Earth's crust since its formation ~4.5 billion years ago. As these isotopes decay, they produce other radioactive isotopes, in turn, leading to the formation of *radioactive decay chains* or *radioactive families*.

By analyzing the decay processes presented before, we can easily notice that, in the *α* decay reactions, the atomic mass of a radionuclide decreases by 4, while in both β decay processes, the atomic mass remains constant. As the atomic mass can only change by decrements of 4, four radioactive chains for the heaviest elements exist that are distinguished by how the atomic mass of their respective isotopes can be divided by 4. Members of one chain cannot decay into isotopes of another chain. All chains end with a stable isotope (usually an isotope of lead).

The four radioactive chains are:

- the *4n* chain (the thorium chain, Figure 9.5), starting with ^{232}Th and ending with ^{208}Pb;
- the *4n+1* chain (the neptunium chain), starting with ^{237}Np and ending with ^{205}Tl;
- the *4n+2* chain (the uranium chain), starting with ^{238}U and ending with ^{206}Pb;
- the *4n+3* chain (the actinium chain), starting with ^{235}U and ending with ^{207}Pb.

In nature, significant amounts of isotopes exist

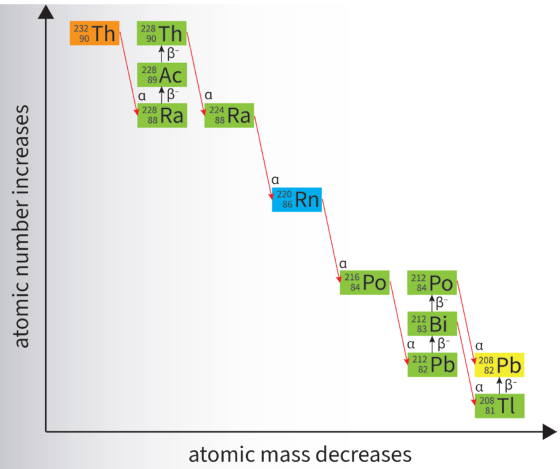


Figure 9.5. The 4n chain (thorium chain). The chain starts with ^{232}Th (orange) and ends with ^{208}Pb (yellow). ^{220}Rn (blue) is a noble gas, all other members are metals. The chain branches out at ^{212}Bi, which has two possible decay paths. α decay is shown by red arrows, $β^-$ decay is shown by black arrows. Note that atomic masses can only decrease by 4.

from three of the four chains. However, most the isotopes of the *4n+1* chain have such small half-lives that they have already decayed into their end products. Significant amounts of only two isotopes can be found in nature from this chain while other members can only be produced artificially in nuclear reactions. The chains that still exist in nature are the major sources of the natural background of ionizing radiation (Figure 9.1). Thus, they are sources of terrestrial radiation, but also produce the two radon isotopes that are the main sources of exposure: ^{222}Rn is produced in the *4n+2* chain, while ^{220}Rn is produced in the *4n* (thorium) chain, hence the name thoron for this latter isotope (Figure 9.5).

4. DOSIMETRY

Dosimetry is the field of science that concerns itself with calculating and measuring the quantity (called *dose*) of ionizing radiation absorbed by matter.

Let us consider the following case (Figure 9.6): a source of ionizing radiation is placed in front of a person. Radiation travels through the air until it reaches the person, after which it will interact with its tissues. Part of that radiation will be absorbed by the tissue, transferring its energy in the process, while another part might pass through.

Several types of doses can be considered for this process, which we will describe below.

4.1. Exposure (X)

We can measure *exposure* in a volume of air in

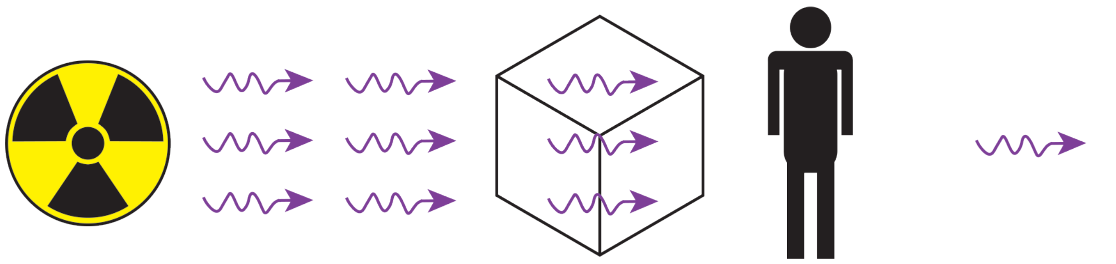


Figure 9.6. Exposure and absorbed dose. A source of a certain activity emits ionizing radiation and a part of it is directed at a target (person). You can measure the exposure in a volume of air in front of the target. From the dose the target was exposed to, a part of its energy will be absorbed by the target. Depending on the type of radiation, not all of its energy might be absorbed.

front of the target. Exposure (X) is defined only for X-rays and γ radiation and is measured as:

$$X = \frac{Q}{m} \quad (9.17)$$

where Q is the total charge of ions produced in a mass of air m.

Thus, X represents the amount of charge (expressed in coulomb) that the electromagnetic ionizing radiation can produce in 1 kg of air. The SI unit for exposure is **C/kg**. An older unit still in use is the **roentgen (R)**, with $1\ R = 2.58 \cdot 10^{-4}$ C/kg.

An *exposure rate* can also be calculated, by expressing how the exposure varies over time. This is measured as units of exposure over units of time: C/(kg·s), R/s, C/(kg·min), etc.

Exposure can be easily measured in air. However, it only informs us of the amount of radiation that reaches the target, but it does not take into account the amount of energy that the radiation will transfer to the target or the biological effects of the exposure. Furthermore, it is only defined for X-rays and γ radiation. For these reasons, at present, other quantities are recommended to be used for expressing the radiation dose.

4.2. The absorbed dose (D)

A quantity that can be defined for all types of ionizing radiation is the *absorbed dose*. According to the International Commission on Radiological Protection (ICRP), the absorbed dose is defined as:

$$D = \frac{d\bar{\varepsilon}}{dm} \quad (9.18)$$

where $d\bar{\varepsilon}$ is the mean energy imparted to matter of mass dm by ionizing radiation.

The SI unit for D is J/kg. However, to distinguish it from other types of doses, when referring to the absorbed dose, a special name is used for the unit J/kg, which is **gray (Gy)**. Thus, for example, a dose of 1 Gy corresponds to a radiation that transfers 1 J/kg of energy to the target.

As for the exposure, we can define an absorbed dose rate by averaging the absorbed dose over time. This is measured in units of dose over units of time: Gy/s, Gy/h, etc.

A legacy unit for the absorbed dose still in use in the US is the **rad** (radiation absorbed dose):

$$1\ rad = 0.01\ Gy \quad (9.19)$$

A related quantity that is also measured in Gy is the Kerma (kinetic energy released per unit mass), which is defined as the sum of the kinetic energies of charged particles liberated (transferred) by uncharged particles per unit mass. This is generally calculated in air (called air-kerma) and can be used instead of exposure:

$$K_{air} = \frac{dE_{tr}}{dm} \quad (9.20)$$

In many situations, the air kerma equals the absorbed dose for uncharged ionizing radiation.

4.3. The equivalent dose (H)

The damaging effect of ionizing radiation to living matter has been observed to depend not only on the absorbed dose, but also on the type of radiation. For example, a tissue that absorbs a dose of 1 Gy of α radiation is much more damaged (approximately 20 times more) than in the case where the same tissue would absorb a dose of 1 Gy of X-rays. Therefore, the absorbed dose is not sufficient, by itself, for describing the biological effect of ionizing radiation.

To account for these differences, each type of ionizing radiation is assigned a radiation weighting factor (w_R) by the ICRP which is based on the relative biological effectiveness (RBE) that

Table 9.2. Values of the radiation weighting factor w_R for different types of radiation[2].

Type of radiation	w_R
EM radiation (X, γ photons)	1
β radiation (electrons, positrons)	1
protons	2
α particles, heavy ions	20
neutrons	calculated depending on their energy using specialized equations

is determined experimentally for the respective radiation. w_R is a number (with no unit of measurement) that shows how much more damaging a particular type of radiation is to living tissue compared to a standard (which is electromagnetic radiation). The currently used w_R values are shown in Table 9.2.

Using the values for the absorbed dose of a particular radiation and taking into account the value of its weighting factor, the **equivalent dose** (called also *dose equivalent*) can be defined as:

$$H = D \cdot w_R \tag{9.21}$$

As w_R is dimensionless (a number with no unit of measurement), the SI unit for H is the same as the SI unit for D, J/kg. For the equivalent dose, we use the special name **sievert (Sv)** for this unit.

Let's consider a few examples:

- 1 Gy of X-rays will correspond to an equivalent dose of 1 Sv;
- 1 Gy of protons will correspond to an equivalent dose of 2 Sv;
- 1 Gy of α radiation will correspond to an equivalent dose of 20 Sv.

The legacy unit for the equivalent dose is the **rem (roentgen equivalent man)**, which is still in use in the US:

$$1\ rem = 0.01\ Sv \tag{9.22}$$

As for the other types of doses, we can define an equivalent dose rate (in Sv/s, Sv/year, rem/s, etc.).

4.4. The effective dose (E)

It has been observed that some tissues are more sensitive to the effects of ionizing radiation, for example, by having an increased risk of cancer development following exposure. In 1906, the Bergonié-Tribondeau law was postulated from experimental observations, stating essentially that **cells are more sensitive to ionizing radiation if they are dividing more rapidly and are less differentiated**. For example, germ cells and embryonic cells are very sensitive to ionizing radiation. In the intervening 100+ years, this law has been shown to have exceptions, however. For example, lymphocytes are highly radiosensitive, even if they are highly differentiated.

In order to account for the variability of ionizing radiation effects on different types of tissue, a tissue weighting factor (w_T) can be established. The w_T is determined by the ICRP based on epidemiological data of cancer incidence following exposure to high doses of ionizing radiation, and is averaged over sex and age. As for w_R, this is a dimensionless quantity. The currently used w_T values are shown in Table 9.3.

Table 9.3. Values of the tissue weighting factor w_T for different types of tissues[2].

Tissue	w_T	Σw_T
Bone marrow, breast, colon, lung, stomach, remainder tissues*	0.12	0.72
Gonads	0.08	0.08
Urinary bladder, oesophagus, liver, thyroid	0.04	0.16
Bone surface, brain, salivary glands, skin	0.01	0.04

* Remainder tissues are: adrenals, extrathoracic regions of the respiratory tract, gall bladder, heart, kidneys, lymphatic nodes, muscle, oral mucosa, pancreas, prostate (male), small intestine, spleen, thymus, and uterus/cervix (female). Each has been assigned a w_T of 0.0086, so that their sum is 0.12.

Using the values of the tissue weighting factors, an **effective dose** can be calculated as:

$$E = \sum_T w_T H_T = \sum_T w_T \sum_R w_R D_{T,R} \tag{9.23}$$

where w_T and w_R are the tissue and radiation weighting factors, respectively, H_T is the equivalent dose received by the respective tissue and $D_{T,R}$ is the absorbed dose for each tissue and type of radiation.

The SI unit for E is **J/Kg** and, as in the case of the equivalent dose, we use the special name **sievert (Sv)** for this unit. Unfortunately, this might lead to confusion, so it is very important to specify each time what quantity is measured when referring to doses in Sv.

Now that you have seen so many types of doses, you might ask which one of these is the most relevant for the medical profession. The truth is, all of them, as they are used in different contexts. The effective dose can be argued to offer the most information regarding the medical risk, but its calculation requires measuring the absorbed dose(s) and calculating the equivalent dose(s) as an intermediate step. However, as the values of w_T have been determined based on exposure of

2 According to the International Commission on Radiological Protection (2007). *The 2007 Recommendations of the International Commission on Radiological Protection. ICRP Publication 103*.

Table 9.4. Physical quantities used in dosimetry and their units of measurement.

Quantity	SI unit	SI unit special name	Legacy unit	Legacy unit conversion
Activity (A)	s^{-1}	Bq (becquerel)	Ci (curie)	1 Ci = $3.7 \cdot 10^{10}$ Bq
Exposure (X)	C/kg	–	R (roentgen)	1 R = $2.58 \cdot 10^{-4}$ C/kg
Absorbed dose (D)	J/kg	Gy (gray)	rad	1 rad = 0.01 Gy
Equivalent dose (H)	J/kg	Sv (sievert)	rem	1 rem = 0.01 Sv
Effective dose (E)	J/kg	Sv (sievert)	rem	1 rem = 0.01 Sv

individuals to high doses of ionizing radiation, the effect of the effective dose at low values can only be estimated based on theoretical models.

4.5. Units of measurement

As you have seen, for each quantity presented in this section, we included two units of measurement: the SI unit and what we called a legacy unit. The ICRP recommends the use of SI units. However, the legacy units were sometimes introduced first and are still commonly used in some parts of the world (for the example, the US). For that reason, we have included both types of units in this chapter. For ease of referring to them, we summarize these units in Table 9.4.

5. INTERACTION OF IONIZING RADIATION WITH LIVING ORGANISMS

5.1. Direct and indirect effects of ionizing radiation

When an organism is exposed to ionizing radiation, two types of effects can be considered at the molecular level:

- *Direct effects*: the damage to the cell or tissue occurs at the level of the molecule that directly absorbs the ionizing radiation;
- *Indirect effects*: the damage to the cell or tissue occurs at the level of a different molecule or molecules than the molecule that absorbed the ionizing radiation.

We call **direct effects** the direct damage resulting from the ionizing radiation being absorbed by a molecule critical for the function of the cell, such as DNA, proteins or lipids. Ionizing radiation can cause single- or, more commonly, double-strand breaks through the direct interaction with DNA. When DNA is damaged, the repair machinery of the cell normally intervenes to repair the damage. If the damage is too extensive, the cell should normally die, usually through controlled death (apoptosis). However, if the damage is light, the cell might survive. If the damage caused by the ionizing radiation was not repaired correctly, the cell might still be able to proliferate and develop into cancer if certain genes are damaged.

The **indirect effects** result from the ionization of other molecules, such as water, that result in the formation of reactive oxygen species or free radicals. These then interact with biologically relevant molecules, damaging them. As water is the main component of the cell, it is the most likely to absorb ionizing radiation following exposure. This is called the *radiolysis of water*. The main ionization event of water can be written as:

$$H_2O \rightarrow H_2O^+ + e^-_{aq} \qquad (9.24)$$

where H_2O^+ is a free radical that contains an unpaired electron (do not mistake it for hydronium, H_3O^+ which is always present in water) and e^-_{aq} is a free electron called a hydrated or a solvated electron.

The two chemical species produced are highly reactive and can, in turn, produce other reactive species (free radicals, reactive oxygen species). We will give only two examples of such reactions:

$$H_2O^+ + H_2O \rightarrow H_3O^+ + HO\cdot \qquad (9.25)$$

$$e^-_{aq} + H_2O \rightarrow H\cdot \ + HO^- \qquad (9.26)$$

where H· and HO· are also highly reactive species (free radicals).

In addition to ionization, absorption of ionizing radiation can also produce excitation of the water molecule, resulting, again, in the production of reactive species. The excitation of water can be written as:

$$H_2O \rightarrow H_2O^* \qquad (9.27)$$

where H_2O^* is an excited water molecule that can, for example, dissociate into free radicals:

$$H_2O^* \rightarrow H\cdot + HO\cdot \qquad (9.28)$$

Overall, the direct and indirect effects of ionizing radiation are summarized in Figure 9.7.

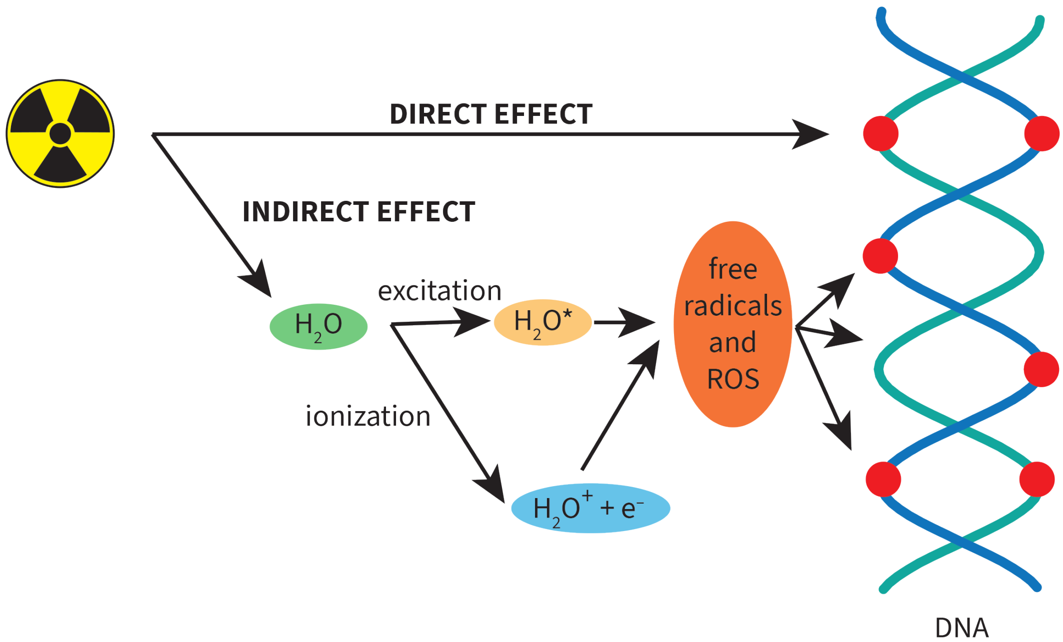


Figure 9.7. Direct and indirect effects of ionizing radiation on the DNA molecule. ROS = reactive oxygen species. Red circles denote areas of DNA damage (single-strand or double-strand breaks).

5.2. Stochastic and deterministic effects of ionizing radiation

Exposure to ionizing radiation can provoke detrimental health effects. Generally, we can classify these effects into two categories, depending on how they have been observed to occur:

- **Stochastic effects** are those that occur randomly (by chance). The probability (but not the severity) of these effects depends on the dose. The main stochastic effect of exposure to ionizing radiation is the development of cancer. Another stochastic example is the appearance of heritable effects (damage to the germ cells that are transmitted to the next generation as genetic mutations).
- **Deterministic effects** are those that are known to occur after a particular dose limit (threshold). Their severity increases with the dose. For example: skin burns, damage to the lens of the eye, acute radiation syndrome.

Stochastic effects can generally be observed when exposure to ionizing radiation is at least moderate. We mentioned that the tissue weighting factors w_T were determined following epidemiological studies on the incidence of cancer in persons exposed to moderate to high doses of ionizing radiation (for example, on survivors of the Hiroshima and Nagasaki bombings). For example, the ICRP lists the chance of developing skin cancer as 0.1 per Sv (0.1 Sv^{-1}). Or, otherwise said, 1000 out of 10000 persons exposed to 1 Sv of ionizing radiation are likely develop skin cancer. At 2 Sv of exposure, 2000 out of 10000 persons are likely to develop skin cancer, and so on. The total risk of cancer for the general population is given by the ICRP as 0.055 Sv^{-1} (550 out of 10000 persons exposed to 1 Sv of ionizing radiation are likely to develop some kind of cancer).

What about low doses? It is hard to experimentally determine the chance of stochastic effects at low doses, due to their relatively low probability. At present, for doses <100 mSv the ICRP uses the linear-no-threshold (LNT) model which states that the incidence of cancer or heritable effects rises proportionally to the equivalent dose (Figure 9.8, left panel). This, is, however, difficult to definitely prove and there are also alternative models. Nevertheless, using the LNT model is the most prudent course of action until another model is universally accepted. As an aside, this is a good place to point out that science (including medicine) many times does not have an "ultimate" answer to a particular question and has to rely on assumptions that are "good enough" or "most likely" until they are proven wrong or replaced by a better model.

An alternative to the LNT model is *radiation hormesis*. This phenomenon has been observed in vitro, in studies on cells. The radiation hormesis model states that low doses of ionizing radiation are actually protective, stimulating cellular repair mechanisms and being generally beneficial.

Hormesis is not acknowledged to occur in humans by regulatory bodies such as the ICRP or NCRP which recommend the LNT model (which is mutually exclusive with hormesis).

Regarding the *deterministic effects* (Figure 9.8, right panel) the science is more clear-cut. At high doses of ionizing radiation, tissue reactions occur that increase in severity as the dose increases.

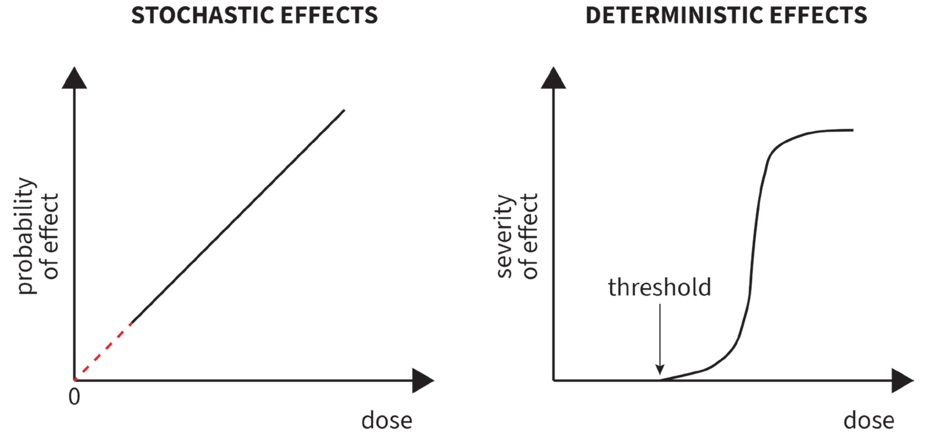


Figure 9.8. Stochastic and deterministic effects of ionizing radiation. Left, according to the LNT model, stochastic effects can appear at any dose, with the probability increasing linearly with the dose. The continuous line shows measured data, while the dashed line shows extrapolated data (at low doses). Right, deterministic effects (tissue reactions) appear after a certain threshold. Many times, the severity of deterministic effects follows a sigmoidal curve, as shown in the figure.

Why does that happen? We have seen that radiation damage is a probabilistic process (there is a chance that a cell irradiated with ionizing radiation will suffer damage). If, out of an entire tissue, only one or several cells are affected, this might have no observable effect on the tissue as a whole (though the chance of future cancer always exists according to the LNT model). As the dose increases, more and more cells are affected until the entire tissue will suffer the effects.

Deterministic effects (tissue reactions) can appear quickly after exposure (hours to weeks) or late after exposure (months to years). Generally, the shorter the exposure time, the lower the threshold is for these effects to appear. For example, damage to the lens of the eye is detected above 0.5 – 2 Gy absorbed in a single event, but if the exposure is split over a larger amount of time, the damage is only observed after 5 Gy. Examples of common deterministic effects and their thresholds for short-term exposures are shown in Table 9.5.

Table 9.5. Examples of deterministic effects and their thresholds[3] for a single, short exposure.

Tissue	Effect	Threshold (Gy)
Testes	Temporary sterility Permanent sterility	0.15 3.5 – 6.0
Ovaries	Sterility	2.5 – 6.0
Lens	Detectable opacities Visual impairment	0.5 – 2.0 5.0
Bone marrow	Depression of hematopoiesis	0.5

5.3. Radiation doses and limits

We have previously stated that we are constantly exposed to natural sources of ionizing radiation and now we are ready to quantify this amount. The NCRP[4] reported that an average individual in the **United States** was exposed in 2006 to **~3.1 mSv/year** from the natural background. The sources of this radiation were shown in Figure 9.1. An estimate for the annual dose received from the **natural background in Romania[5] was of ~2.7 mSv/year** in 2000, out of which ~1.5 mSv were from ^{222}Rn and ^{220}Rn exposure.

At present, the **annual dose limit** for exposure to **artificial sources of ionizing radiation** recommended by the ICRP is **1 mSv/year for a member of the general population** (given as equivalent dose). **This does not include exposure to medical procedures**, which should be evaluated on a case-by-case basis.

For workers whose activity occurs in environments where they can be exposed to ionizing radiation, the **occupational limit** was set at **20 mSv/year averaged over a period of 5 years, and no more than 50 mSv in a single year** (given as equivalent dose).

What are actually dangerous doses? Doses above 100 mSv are clearly correlated with cancer risk, as assessed from the previously mentioned studies on Japanese atomic bomb survivors. At doses between 10 and 100 mSv, the data are insufficient to draw a clear conclusion on the dose dependence, and are still disputed. However,

3 According to the International Commission on Radiological Protection (2007). *The 2007 Recommendations of the International Commission on Radiological Protection. ICRP Publication 103*.

4 National Council on Radiation Protection & Measurements. (2009). *NCRP Report No. 160, Ionizing Radiation Exposure of the Population of the United States*.

5 Iacob, O., & Botezatu, E. (2000). Exposures from Natural Background Radiation in Romania. *Bulgarian Journal of Physics*, 27(3), 98-101.

Table 9.6. Examples of radiation doses for various medical imaging procedures that use ionizing radiation[9]. The rightmost column gives the time it would take for a similar dose to be imparted from the natural background of ionizing radiation in Germany, which was estimated at 2.1 mSv/year. Doses higher than a year's exposure to the natural background are shown in bold for emphasis. PA = postero-anterior. AP = antero-posterior.

Imaging procedure	Effective dose (mSv)	Equivalent to how many PA chest X-rays	Time it would take to receive the dose from the natural background
X-ray (limbs)	< 0.01	< 0.5	< 2 days
X-ray (AP/PA cranium)	0.021	1.2	4 days
X-ray (chest)	0.018	1	3 days
X-ray (AP/PA abdomen)	0.34	20	2 months
Mammography	0.36	20	2 months
Coronary angiography	3.2	180	**1.5 years**
Arteriography leg and pelvis	4.5	250	**2.2 years**
Endovascular aneurysm repair (EVAR) of the aorta	17	950	**8 years**
CT (brain)	1.6	90	9 months
CT (chest)	5.1	280	**2.4 years**
CT (abdomen and pelvis)	11	600	**5.2 years**
CT angiography (aorta)	10	550	**4.8 years**
Cerebral scintigraphy	4.5	250	**2.1 years**
Thyroid scintigraphy	0.9	50	5 months
PET scan	4.6	250	**2.2 years**

as we have seen, the LNT model considers that cancer risk exists even at very low doses. Thus, according to the LNT model, there is no "safe" dose of ionizing radiation and even moderate doses such as those commonly employed in Computed Tomography (CT) scans can potentially lead to development of cancer.

What about lethal doses (LD)? The ICRP[6] lists the dose at which 50% of the persons exposed die after 60 days as $\mathbf{LD_{50/60} = 4\ Sv}$. Death generally results from hematopoietic failure. Survival odds can be improved using supportive medical treatment. At doses above 5 Sv, severe gastrointestinal damage appears and the risk of death is even higher. The dose at which 90% of persons exposed die was estimated to be $\mathbf{LD_{90} = 5 - 7\ Sv}$. Note that all these doses correspond to whole body exposure over a short period of time.

5.4. Doses in medical procedures

Let us now consider the doses that a patient receives in some common medical procedures. We will not ask you to memorize these doses! The medical sources of ionizing radiation are imaging procedures using ionizing radiation (most commonly radiography and CT scans) and radiotherapy. Overall, the ionizing radiation dose from medical procedures was estimated to be 2.16 mSv/year for persons in the US[7] in 2016, while the EU[8] reported an average of 1.66 mSv/year in 2015, with Romania having the lowest doses in the EU, at 0.343 mSv/year.

Table 9.6 shows some examples of doses that a patient is subjected to in different medical imaging procedures. For ease of better contextualizing the doses, the doses are compared to the dose for a single chest radiography. Additionally, the time that it would take for the natural background to impart the same effective dose is also given in the table.

You can observe that a lot of these procedures impart, in a few minutes or tens of minutes, doses that the patient might naturally receive over months to years of exposure to the natural background. While a single X-ray image might be harmless, CT scans subject the patient to doses equivalent to hundreds of individual X-rays. This

6 According to the International Commission on Radiological Protection (2007). *The 2007 Recommendations of the International Commission on Radiological Protection. ICRP Publication 103*.

7 National Council on Radiation Protection & Measurements. (2019). *Report No. 184 – Medical Radiation Exposure of Patients in the United States*.

8 Directorate-General for Energy of the European Commission. (2015). *RP 180 Medical Radiation Exposure of the European Population, Part 1 & Part 2*.

9 According to the German Commission on Radiological Protection (SSK). (2019). *Recommendations for medical imaging procedures.* Adopted at the 300th SSK meeting on 27 June 2019.

might cause the patient to develop cancer years later, especially if several procedures are performed over a short amount of time.

Let's give two examples:

▸ Endovascular aneurysm repair (EVAR) is a fluoroscopic procedure that exposes patients to doses of 10 – 100 mSv and has been shown to increase the risk of abdominal cancer in a study[10] that analyzed patients over 50 years old subjected to EVAR in England;

▸ Radiotherapy employs ionizing radiation to destroy cancerous tumors. Typical doses used in a session of radiotherapy vary in the range 1.5 – 3 Gy and are applied to the site of the tumor in several fractions, with the cumulative doses ending up at tens of Gy. Radiotherapy has a clear risk of developing secondary cancers (usually in organs adjacent to the site of the primary cancer that were exposed to the radiation).

Currently, the ICRP does not impose a limit to the doses of ionizing radiation received from medical procedures, but advises that proper justification should be had for procedures that involve the use of ionizing radiation on patients. Thus, three levels are recommended for justification:

▸ that the use of the procedure in general does more good than harm (this is assumed to be true);

▸ that the procedure employed is going to improve the diagnosis of the patient;

▸ that the proposed procedure is justified for the individual patient.

5.5. Types of exposure. Biological and effective half-life

When considering exposure to ionizing radiation, the position of the radiation source relative to the patient also has to be considered. Thus, irradiation can be:

▸ *External*, if the source is outside the body. This is usually the case in most imaging procedures as well as for most radiotherapy procedures;

▸ *Internal*, when the irradiation source is inside the body. This occurs, for instance, following nuclear fallout, but it is also a part of some imaging or radiotherapy techniques.

If considered in terms of radiation protection, internal irradiation has the potential of being more harmful as a general rule, as it is more likely to affect the function of internal organs. When the radiation source is external, the skin serves as a natural barrier for some types of radiation, as we will see below.

Table 9.7. Half-lives of some radioisotopes resulting from nuclear fallout.

Radioisotope	$t_{1/2,phys}$	$t_{1/2,biol}$	$t_{1/2,eff}$
^{90}Sr	28 years	~40 years*	16 years
^{131}I	8 days	80 days	7 days
^{137}Cs	30 years	70 days	69 days

* various values are reported in literature, from 10-40 years.

When radioisotopes are present inside the body, the time that it will take for them to leave the body has to be considered. We have seen that each radioisotope has a specific half-life that shows how long it takes for the number of its nuclei to reduce to half. We will call this the *physical half-life* in order to distinguish it from the *biological half-life*, which is the time for half of the amount of a particular radionuclide to be eliminated from the body through biological processes.

To account for the disappearance of a radionuclide through both physical (radioactive decay) and biological processes, an *effective half-life* can also be calculated by taking into account both the physical and biological half-life:

$$\frac{1}{t_{1/2,eff}} = \frac{1}{t_{1/2,phys}} + \frac{1}{t_{1/2,biol}} \tag{9.29}$$

where $t_{1/2,eff}$ is the effective half-life, $t_{1/2,phys}$ is the physical half-life and $t_{1/2,biol}$ is a biological half-life of a particular radionuclide.

As a general rule, the effective half-life is always smaller or at most roughly equal to the smallest of the physical and biological half-lives. Table 9.7 shows the physical, biological and effective half-lives of some radioisotopes commonly produced in nuclear fallout.

6. PROTECTION FROM IONIZING RADIATION

6.1. Detection of ionizing radiation

Our senses are not able to detect ionizing radiation. Symptoms appear in the body only following exposure to very high doses (hundreds of mSv at least). Thus, accurate detection of ionizing radiation has to be performed using specialized instruments. Several types of instruments are available for the detection of ionizing radiation.

These can be classified as:

▸ Photographic film: used in personal detectors as well as in older equipment for radiography. More on photographic film in radiography is described in the chapter on Medical imaging;

▸ Gas-filled detectors: contain an inert gas that is ionized by the passage of the radiation,

10 Markar, S. R., Vidal-Diez, A., Sounderajah, V., Mackenzie, H., Hanna, G. B., Thompson, M., . . . Karthikesalingam, A. (2019). A population-based cohort study examining the risk of abdominal cancer after endovascular abdominal aortic aneurysm repair. *J Vasc Surg*, 69(6), 1776-1785 e1772. doi:*10.1016/j.jvs.2018.09.058*.

the produced charges are detected electrically afterwards. Examples are: ionization chambers, proportional counters, the Geiger-Müller counter;
▸ Scintillation detectors: contain a material that emits UV or visible radiation when it absorbs ionizing radiation. An example, the gamma camera, is described in the chapter on Medical imaging;
▸ Solid-state detectors: contain a semiconductor material in which ionizing radiation can produce charge movements that can be detected electrically. They are more sensitive than gas-filled detectors.

6.2. Penetration power of ionizing radiation. The Bragg curve

The *penetration power* is different for the various types of ionizing radiation. Particles that are large and carry electrical charge generally travel less deep inside matter, as they are quickly absorbed. Therefore:

▸ α particles quickly transfer all their energy and are absorbed by even a thin sheet of paper (Figure 9.9). If a patient is exposed to α radiation, the damage (on areas not covered by clothes) will be limited to the skin and cornea. α radiation is highly dangerous, however, when the source of exposure is internal. Because of its low power of penetration, α radiation is not used in medical imaging;
▸ β radiation is more penetrating than α particles due to its lower mass and charge. β radiation can be blocked by shields made out of light metal or plastic;
▸ Due to its zero charge and zero resting mass, electromagnetic radiation is the most penetrating. Thus, X-rays or γ radiation can only be attenuated (their intensity is reduced) by thick lead shields;
▸ For neutrons, their penetration power depends on their energy. Multiple successive shields have to be used in order to first slow down and then stop neutrons.

The penetrating power of radiation is correlated with the transfer of energy to the matter that it passes through. This is calculated in dosimetry

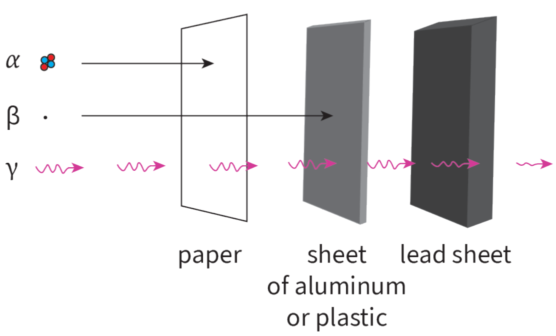


Figure 9.9. Penetrative power of different types of ionizing radiation. Description is in the text.

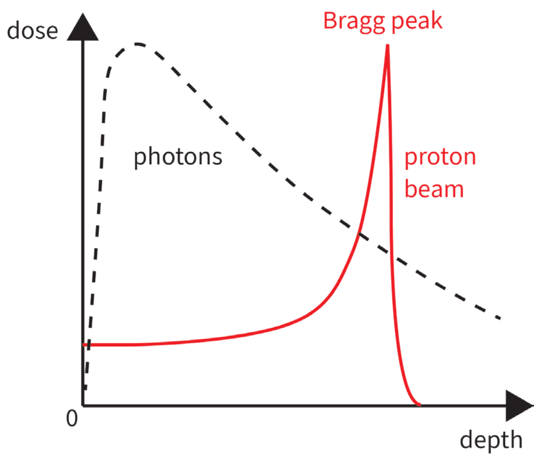


Figure 9.10. The Bragg curve for protons (solid line) and photons (dashed line). Photons deposit much of their energy at a lower depth and then slowly deposit the rest of their energy as they travel through the tissue. Protons deposit most of their energy at a higher depth, at the Bragg peak.

through the LET (linear energy transfer):

$$LET = \frac{dE}{dl} \tag{9.30}$$

where E is the energy imparted by the radiation to the medium by the radiation and l is the distance travelled by the particle.

α and β particles, protons and neutrons are considered high LET radiation while X-rays and γ radiation are considered low LET.

As particle ionizing radiation travels through matter, particles lose more and more of their energy until they are stopped. For large charged particles (α particles, protons), this occurs suddenly, with them depositing all their energy through the form of a *Bragg peak* (Figure 9.10).

Knowledge of the absorption profile for different types of radiation is essential in radiotherapy, where the aim is to deliver as much of the energy of the ionizing radiation as possible to the tumor, while limiting the dose received by the surrounding healthy tissue. The position of the Bragg peak of protons depends on their energy, an effect which can be used to concentrate the proton beam on the tumor site. By comparison, radiotherapy using photons exposes healthy tissue to ionizing radiation a lot more that a proton beam.

6.3. Physical protection

Three principles of radiation safety should always be kept in mind: **time**, **distance** and **shielding**:
▸ The best method of protection from ionizing radiation is reducing the **time** of exposure as much as possible;
▸ When exposure cannot be avoided, care should be taken to increase the **distance** to the source of radiation;

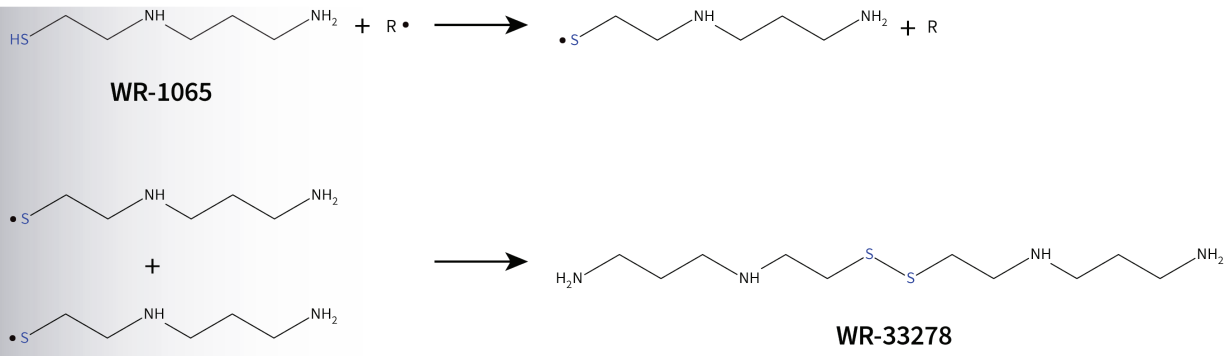


Figure 9.11. Radioprotection activity of WR-1065. WR-1065 reacts with a free radical, turning into a thiol radical (top reaction). Subsequently, two WR-1065 radicals can form a dimer, named WR-33278 (bottom reaction).

▸ Finally, adequate protective **shielding** should be worn that depends on the type and penetrative power of the radiation (Figure 9.9).

6.4. Chemical protection

Limited protection can be provided against indirect effects of ionizing radiation using medicine. Generally, **cells are more sensitive to the indirect effects of ionizing radiation if they have a higher oxygen concentration and a higher amount of water present.**

Compounds that can be used in radioprotection can be classified as:

▸ radiation mitigators: substances that are used after irradiation to reduce the effects;

▸ radioprotectors: drugs that can be administered before the irradiation.

Various mechanisms can be used by radioprotectors, of which we will list a few: scavenging free radicals and reactive oxygen species, formation of disulfide compounds, inhibition of oxidation, absorption of secondary UV radiation, inhibition of metabolism, etc.

An example of protective reaction is the scavenging of free radicals performed by thiol compounds. At the moment, only two radioprotective agents are approved by the US Food and Drug Administration (FDA): amifostine, and palifermin. Palifermin is a keratinocyte growth factor that is used to protect from oral mucositis following radiotherapy. Amifostine is a prodrug that is transformed in the body by alkaline phosphatase into its active metabolite, WR-1065. WR-1065 contains a thiol group that can scavenge free radicals according to the mechanism in Figure 9.11.

A different example of chemical protection can be provided to the thyroid in case of ^{131}I internal exposure. Accumulation of ^{131}I in the thyroid can be prevented by administering iodide or iodate salts to patients exposed to ^{131}I. This saturates the thyroid, that normally absorbs iodine, with non-radioactive iodine, thereby protecting it from ionizing radiation.

REFERENCES

Ahmad, M. I., Ab Rahim, M. H., Nordin, R., Mohamed, F., Abu-Samah, A., & Abdullah, N. F. (2021). Ionizing Radiation Monitoring Technology at the Verge of Internet of Things. *Sensors (Basel), 21*(22). doi:10.3390/s21227629

Aliper, A. M., Bozdaganyan, M. E., Sarkisova, V. A., Veviorsky, A. P., Ozerov, I. V., Orekhov, P. S., . . . Osipov, A. N. (2020). Radioprotectors.org: an open database of known and predicted radioprotectors. *Aging (Albany NY), 12*(15), 15741-15755. doi:10.18632/aging.103815

Allen, C. M. (2022). *Digital Radiographic Exposure: Principles & Practice*. Retrieved from https://umsystem.pressbooks.pub/digitalradiographicexposure/

Baeyens, A., Abrantes, A. M., Ahire, V., Ainsbury, E. A., Baatout, S., Baselet, B., . . . Wozny, A.-S. (2023). Basic Concepts of Radiation Biology. In S. Baatout (Ed.), *Radiobiology Textbook* (pp. 25-81). Cham: Springer International Publishing.

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Berrington de Gonzalez, A., Mahesh, M., Kim, K. P., Bhargavan, M., Lewis, R., Mettler, F., & Land, C. (2009). Projected cancer risks from computed tomographic scans performed in the United States in 2007. *Arch Intern Med, 169*(22), 2071-2077. doi:10.1001/archinternmed.2009.440

Borrego-Soto, G., Ortiz-Lopez, R., & Rojas-Martinez, A. (2015). Ionizing radiation-induced DNA injury and damage detection in patients with breast cancer. *Genet Mol Biol, 38*(4), 420-432. doi:10.1590/S1415-475738420150019

Directorate-General for Energy of the European Commission. (2015). *RP 180 Medical Radiation Exposure of the European Population, Part 1 & Part 2*. Retrieved from https://energy.ec.europa.eu/publications/rp-180-medical-radiation-exposure-european-population-part-1-part-2_en

Dracham, C. B., Shankar, A., & Madan, R. (2018). Radiation induced secondary malignancies: a review article. *Radiat Oncol J, 36*(2), 85-94. doi:10.3857/roj.2018.00290

Durante, M., & Loeffler, J. S. (2010). Charged particles in radiation oncology. *Nat Rev Clin Oncol, 7*(1), 37-43. doi:10.1038/nrclinonc.2009.183

Franklin, K., Muir, P., Scott, T., & Yates, P. (2019). *Introduction to Biological Physics for the Health and Life Sciences*: Wiley.

German Commission on Radiological Protection (SSK). (2019). *Recommendations for medical imaging procedures. Adopted at the 300th SSK meeting on 27 June 2019*. Retrieved from https://www.ssk.de/SharedDocs/Beratungsergebnisse_E/2019/2019-06-27Orientie_e.html

Harrison, J. D., Balonov, M., Martin, C. J., Ortiz Lopez, P., Menzel, H. G., Simmonds, J. R., . . . Wakeford, R. (2016). Use of effective dose. *Ann ICRP, 45*(1 Suppl), 215-224. doi:10.1177/0146645316634566

Iacob, O., & Botezatu, E. (2000). Exposures from Natural Background Radiation in Romania. *Bulgarian Journal of Physics, 27*(3), 98-101.

International Commission on Radiological Protection. (2007). The 2007 Recommendations of the International Commission on Radiological Protection. ICRP Publication 103.

Kardamakis, D., Baatout, S., Bourguignon, M., Foray, N., & Socol, Y. (2023). History of Radiation Biology. In S. Baatout (Ed.), *Radiobiology Textbook* (pp. 1-24). Cham: Springer International Publishing.

Kim, I., & He, Y.-Y. (2014). Ultraviolet radiation-induced non-melanoma skin cancer: Regulation of DNA damage repair and inflammation. *Genes & Diseases, 1*(2), 188-198. doi:https://doi.org/10.1016/j.gendis.2014.08.005

Lin, E. C. (2010). Radiation risk from medical imaging. *Mayo Clin Proc, 85*(12), 1142-1146; quiz 1146. doi:10.4065/mcp.2010.0260

Markar, S. R., Vidal-Diez, A., Sounderajah, V., Mackenzie, H., Hanna, G. B., Thompson, M., . . . Karthikesalingam, A. (2019). A population-based cohort study examining the risk of abdominal cancer after endovascular abdominal aortic aneurysm repair. *J Vasc Surg, 69*(6), 1776-1785 e1772. doi:10.1016/j.jvs.2018.09.058

National Council on Radiation Protection & Measurements. (2009). *NCRP Report No. 160, Ionizing Radiation Exposure of the Population of the United States*. Retrieved from https://ncrponline.org/wp-content/themes/ncrp/PDFs/ExecSumm_NCRP-Report-No-160.pdf

National Council on Radiation Protection & Measurements. (2019). *Report No. 184 – Medical Radiation Exposure of Patients in the United States*. Retrieved from https://ncrponline.org/wp-content/themes/ncrp/PDFs/Product-attachments/184/overview.pdf

Petoussi-Henss, N., Satoh, D., Endo, A., Eckerman, K. F., Bolch, W. E., Hunt, J., . . . Yoo, S. J. (2020). ICRP Publication 144: Dose Coefficients for External Exposures to Environmental Sources. *Annals of the ICRP, 49*(2), 11-145. doi:10.1177/0146645320906277

Pierce, D. A., & Preston, D. L. (2000). Radiation-related cancer risks at low doses among atomic bomb survivors. *Radiat Res, 154*(2), 178-186. doi:10.1667/0033-7587(2000)154[0178:rrcral]2.0.co;2

Widmann, G., Beyer, A., Jaschke, W., Luger, A., Zoller, H., Tilg, H., . . . Verius, M. (2023). Identification and characterization of patients being exposed to computed-tomography associated radiation-doses above 100 mSv in a real-life setting. *Eur J Radiol Open, 10*, 100470. doi:10.1016/j.ejro.2022.100470

CHAPTER 10

BIOPHYSICS OF VISION

Prerequisite knowledge

- Interaction of light with matter. Chromophores
- Refraction of light, refractive index
- Types of lenses, refractive power, focal length, types of images
- Resting membrane potential, depolarization, hyperpolarization
- Cis-trans isomerism

1. THE PROCESS OF VISION

If somebody asked you which of your senses you value the most, you would probably say the sense of sight, because it allows you to gain more information from the surrounding medium than you do with any of the other senses. We use our eyes for almost any activity we perform, as they enable us to perceive the shape, color, size and movement of the objects around us by detecting and analyzing the light[1] they reflect or emit. The eye is able to detect various light intensities and even dim light, but it cannot sense objects in the absence of light.

The sensation of sight requires a succession of complex processes, including:

- Refraction of light rays by the transparent media and their focus in the posterior part of the eye (on the retina). As a result, a *real, upside-down image* is formed on the retina;
- Absorption and conversion of light energy into chemical and electrical signals performed by light-sensitive cells in the retina (photoreceptors);
- Parallel processing of various characteristics of visual information such as color, shape, contrast, dimension and movement, which starts in the retina and continues all the way to and within the visual cortex;
- Integration of these distinct aspects to produce the conscious sensation of sight. Note that the sensation of vision results from the spatial and temporal electrical activity of the cortical cells, and does not represent a 1:1 reproduction of the physical image formed on the retina.

The main elements of the visual system are: the eyes, afferent and efferent neural pathways, intermediate processing structures and cortical areas in the occipital lobe.

In this chapter, only the structures of eye and their role in the process of vision will be addressed.

2. THE STRUCTURE OF THE EYE

Figure 10.1 shows a cross-section through the human eye, with the main components labeled. The eye is approximately spherical in shape, with a diameter of about 2.5 cm. It is a hollow organ (filled with fluids) and its wall is composed of three main layers of tissue (also called tunics):

- *The fibrous layer* (the outermost tunic) contains the **sclera** (or white of the eye), continuing in the anterior part with the **cornea,** which is transparent.

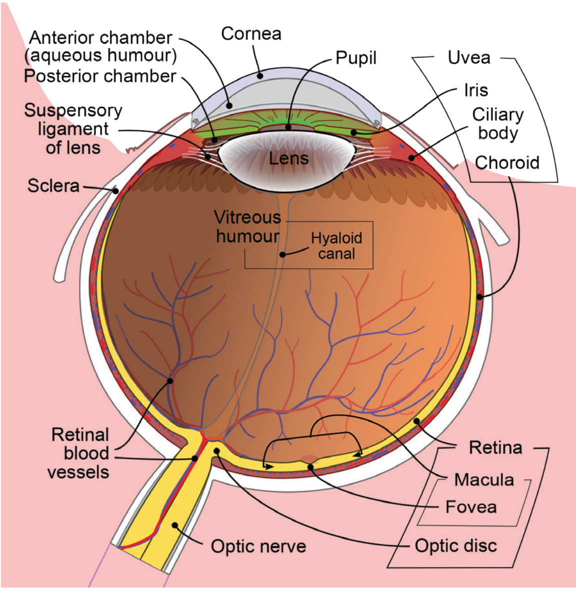


Figure 10.1. Anatomy of the eye.[2]

1 The term “light” denotes here the visible domain of electromagnetic radiation (roughly between 400 – 760 nm), as described in the chapter on Photobiology.

2 Image by Rhcastilhos and Jmarchn (*https://upload.wikimedia.org/wikipedia/commons/1/1e/Schematic_diagram_of_the_human_eye_en.svg*) available under a Creative Commons license (*https://creativecommons.org/licenses/by-sa/3.0/deed.en*).

▹ The sclera is a dense connective tissue (containing collagen and elastin fibers), thickest over the posterior surface of the eye and thinnest over the anterior surface. It is important for support and protection of the eye and it also serves as an attachment site for the tendons of the extraocular muscles;

▹ The cornea is the first ocular medium crossed by light and, due to its curvature and refractive index, it aids in the focusing process.

▸ *The vascular layer* (the middle tunic, or uvea) consists of three continuous structures: the **choroid**, the **ciliary body**, and the **iris** (the colored part of the eye).

▹ The choroid contains numerous blood vessels, lymphatic vessels and is essential for nourishing the retina. Also, it contains a great amount of the pigment melanin, important for limiting the reflection of light within the eye which might lead to unclear images.

▹ Toward the anterior part, there is the ciliary body, a ring-like thickened region budging into the interior of the eye, surrounding the lens of the eye. It has a muscular component, the **ciliary muscles** and a vascular component, the **ciliary processes** (folds of epithelium over the ciliary muscles). **Suspensory ligaments** are connective tissue fibers attached to the tips of the ciliary processes and which hold the lens in place (Figure 10.2);

▹ The iris is another ring-like structure composed of pigmented cells, blood vessels and two types of smooth muscle fibers with opposite actions which, by their contraction, change the diameter of the **pupil** (the opening in the middle), thus regulating the amount of light that enters the eye. Due to the ciliary muscles, the position of the lens is centered on the pupil, so that light that passes through it will also pass through the lens. The color that you see when looking into somebody's eyes depends on the amount of the pigment melanin present in the front layer of the iris.

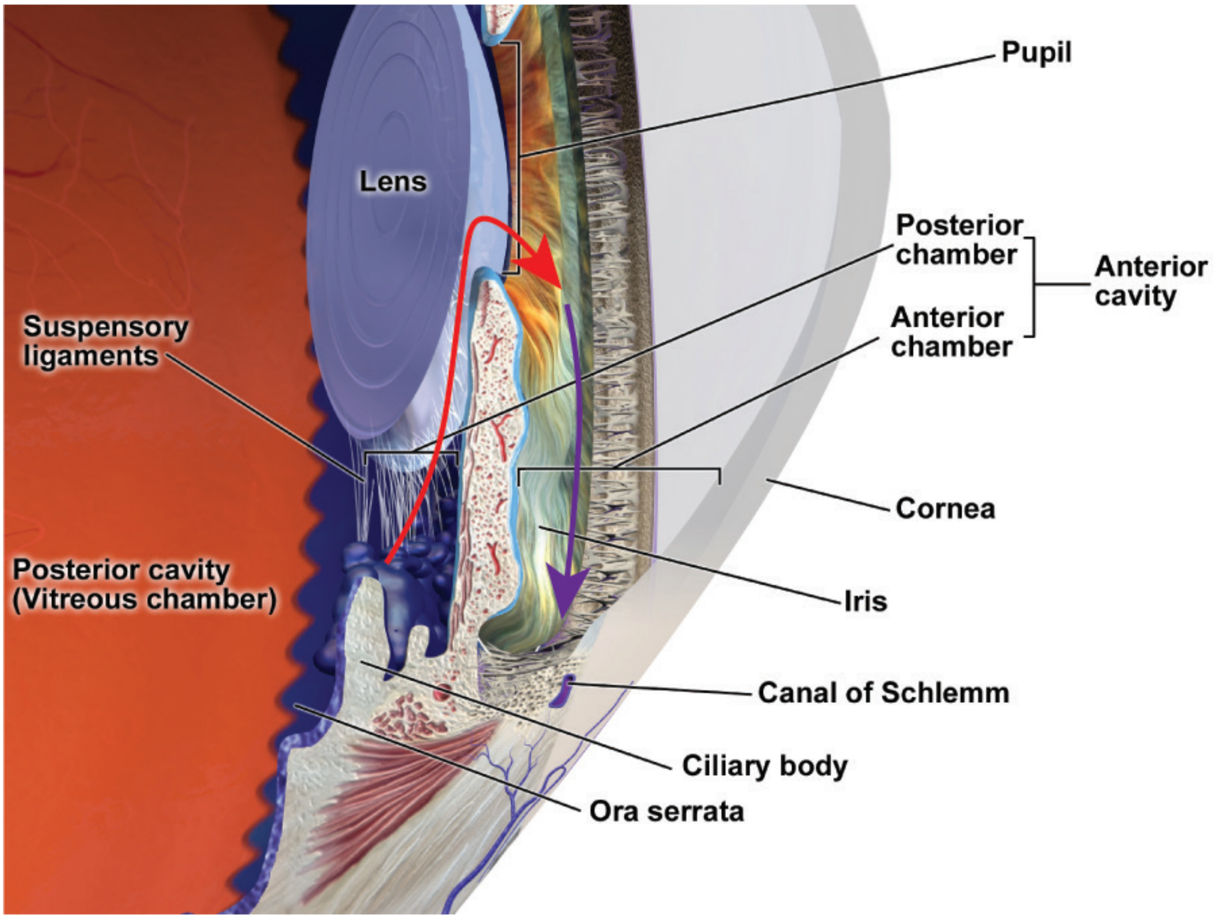


Figure 10.2. Cross-section through the eye[3], showing from anterior to posterior: the cornea, the anterior chamber, the posterior chamber, the lens and part of the posterior cavity. The aqueous humor is secreted by the epithelium of the ciliary body into the posterior chamber and it passes into the anterior chamber through the pupil (red arrow). From the anterior chamber it enters the venous circulation through Schlemm's canal (purple arrow) or diffuses through the ciliary muscle and adjacent structures.

3 Modified from a copyright-free image provided by Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". WikiJournal of Medicine 1 (2). DOI:*10.15347/wjm/2014.010*. ISSN 2002-4436.

People with dark eyes have a high melanin content, while people with blue eyes have less melanin. There is no blue pigment in the eye, you get the sensation of blue color due to the selective absorption and scattering of light by the structures of the iris.

- *The nervous layer* (the innermost tunic) or **retina** is the site of the phototransduction process, which will be detailed later. A retinal area of great importance is the **macula lutea**, with a shallow pit in the center of it, the **fovea**. The **optic disc** is a retinal region free of photoreceptors which serves as the exit point for the neurons of the optic nerve (cranial nerve II) and veins and as the entry point for arteries.

The ciliary body and the lens divide the interior eye into two fluid filled cavities:

- the *anterior cavity*: small, filled with a watery fluid called **aqueous humor**;
- the *posterior cavity*: large, filled with a gel-like fluid called **vitreous humor**, also called *vitreous body*.

The anterior cavity is further divided by the iris into the **anterior chamber** and the **posterior chamber**.

The **aqueous humor** is a transparent fluid similar in composition to blood plasma, but lacking proteins. Its main roles are providing nutrients to the cornea and lens and to maintain intraocular pressure.

The fluid pressure within the eye (*intraocular pressure*) helps to maintain the eye's shape. It can be measured using a method called tonometry and values between 11 – 21 mmHg are considered normal (with some variations during the day) for the anterior chamber. The amount of aqueous humor present at a certain moment in the anterior cavity (and thus producing the intraocular pressure) is the net result of two processes: secretion and outflow of the aqueous humor.

The aqueous humor is secreted by the ciliary epithelium into the posterior chamber of the anterior cavity of the eye. From here, the aqueous humor flows between the posterior part of the iris and the anterior part of the lens and finally into the anterior chamber through the pupil. The outflow of the aqueous humor from the anterior chamber can happen in two main ways:

- The *conventional outflow pathway* (which accounts for the vast majority of the aqueous humor outflow): by drainage of the trabecular meshwork (a group of tiny canals located at the angle between the iris and the cornea), **Schlemm's canal** (Figure 10.2), other small channels and finally entering the venous circulation.
- The *unconventional outflow pathway*: a small amount of aqueous humor enters the venous circulation by flowing between ciliary muscle bundles, suprachoroidal and scleral tissues.

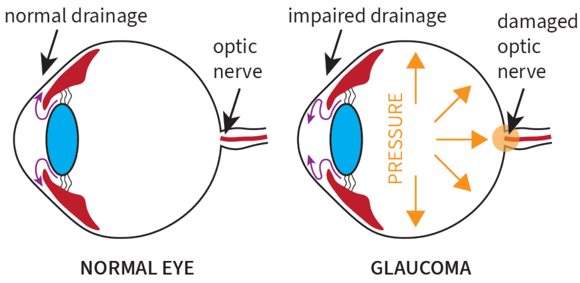


Figure 10.3. Eye with glaucoma. In the normal eye (left panel) drainage of the aqueous humor occurs normally through the trabecular meshwork into Schlemm's canal. In glaucoma (right panel), drainange is impaired, leading to an increase of the intraocular pressure, resulting in damage to the ganglion cells of the optic nerve.

Also, there is yet another way through which the aqueous humor leaves the anterior chamber: by returning into the posterior chamber through the iris.

The production of aqueous humor also keeps the vitreous body at a certain pressure. An imbalance between the secretion and outflow of the aqueous humor leads to a higher-than-normal pressure in the anterior chamber, a pathological condition called **glaucoma** which, if left untreated, can lead to blindness. A common cause for glaucoma is the blockage of Schlemm's canal which leads to an excess of fluid in the anterior chamber because of which the intraocular pressure increases. As a consequence, the force exerted by the vitreous humor onto the surface of the retina will be stronger and potentially damaging for the blood vessels and the axons of the optic nerve (Figure 10.3).

The **vitreous humor** is a gel-like transparent fluid making up most of the eye's volume. It is composed mainly of water, but unlike the aqueous humor, it contains a network of collagen fibers. The vitreous humor plays roles in refraction, eye development and intraocular oxygen metabolism.

3. THE EYE AS AN OPTICAL SYSTEM

3.1. Transparent media of the eye

When we look around, we see various objects because they emit light, or, in most cases, because they reflect the light coming from a light source and at least some of those rays of light enter our eyes. A condition that must be fulfilled in order for us to see clear images of objects is that light must cross several transparent media in our eyes in order to reach the surface of the retina. The eye's transparent media crossed by light, from

Table 10.1. Average values for indices of refraction of air and the transparent media in the relaxed eye, according to the American Academy of Ophthalmology[4]. Note that the highest difference between indices of refraction of two adjacent media is at the air-cornea interface.

Optical medium	Refractive Index
Air	1
Cornea	1.3765
Aqueous humor	1.3335
Crystalline lens	1.4 – 1.42
Vitreous humor	1.335

anterior to posterior, are:

- The **cornea**. It is an avascular and transparent tissue whose thickness varies between 0.5 mm (at the center) and 1.2 mm (at the periphery), containing mainly a dense network of highly ordered collagen fibers. Due to its shape, it acts as a *positive meniscus*;
- The aqueous humor;
- The **crystalline lens** (also simply called "the lens"). It is a *biconvex lens* (with curvature radii R_1 = 10 mm and R_2 = 6 mm) composed of closely packed cells arranged in concentric shells in an onion-like manner and containing mainly water and transparent proteins. These cells have a high content of proteins called α-crystallins (hence the name of crystalline lens), contributing to the refractive index of the lens which increases from the periphery toward the center. The lens allows light to pass through (is transparent), having the primary function of focusing the light rays on the surface of the retina. In addition, the lens (like the cornea) acts as a filter that absorbs part of the electromagnetic radiation with short wavelengths (UV, X, γ), thus playing a role in the protection of the retina. The power of the lens is variable due to the change in curvature during the accommodation process (see below);
- The vitreous humor.

Under normal conditions, each of these media behaves like a spherical converging lens and together they form an optical system refracting the rays of light so that images of objects are focused on the retina.

You were shown in the Biophysics practical sessions and seminars some simple rules that describe how to trace the path of light rays passing through a lens in order to find the position and the type of image it projects. If we want to do the same for the optical media of the eye, we must analyze the refraction at each of the four interfaces: two at the cornea (air-cornea and cornea-aqueous humor) and two at the lens (aqueous humor-lens and lens-vitreous humor) also taking into account the refractive index of each medium (Table 10.1).

3.2. General notions about lenses

A brief reminder about lenses is provided below (these notions are discussed in more detail in the Biophysics practical sessions and seminars):

The **optical power** of a lens, *P*, is the inverse of the focal length, *f*. It depends on the index of refraction of the lens, n_l, that of the medium in which the lens is placed in, n_m, and also on the radii of curvature of the two surfaces, R_1 and R_2 (Figure 10.4) The relation is described by the so called *lensmaker's equation*, which, for a thin lens, has the following form:

$$P = \frac{1}{f} = \frac{(n_l - n_m)}{n_m}\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{10.1}$$

The above equation is valid for both converging and diverging lenses if certain sign conventions are used. For example, if light comes from left to right, the horizontal distances to the right of the center of the lens are positive (e.g. R_1 in Figure 10.4), while those to the left are considered to be negative (e.g. R_2 in Figure 10.4). As a result, the power of a converging lens is positive, while the power of a diverging lens is negative. Power is measured in **diopters** (D), with 1 D = 1 m^{-1}.

When two or more lenses have the same

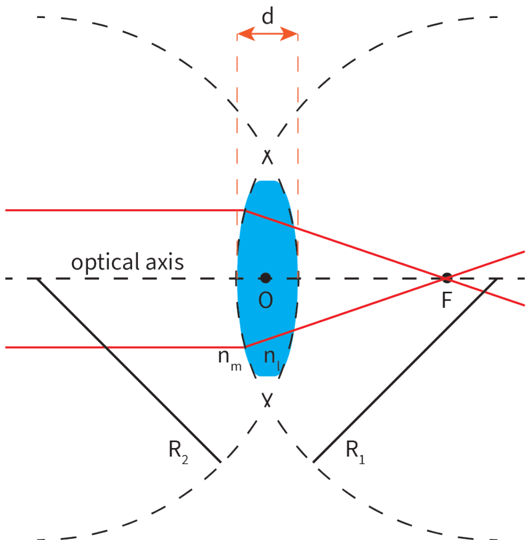


Figure 10.4. Schematic representation of a biconvex lens. A lens is considered thin if its thickness, *d*, is much smaller than the absolute value of its radii of curvature, R_1 and R_2. Two rays of light coming from the left and passing through a thin lens parallel to its principal axis will be refracted at each surface so that they will intersect the principal axis on the other side of the lens (right side) in a point on the principal axis called the focal point (F). The distance between the center of the lens (O) and the focal point is the focal length, *f*. n_l and n_m denote the refractive indices of the lens and medium, respectively.

4 Daniel Palanker, Optical Properties of the Eye, *https://www.aao.org/education/munnerlyn-laser-surgery-center/optical-properties-of-eye*

principal axis, they form an *optical system*. If the lenses are in direct contact, the overall power of the optical system, P_{system}, is the sum of their individual powers, P_{lens1}, P_{lens2}, etc.:

$$P_{system} = P_{lens1} + P_{lens2} + P_{lens3} + \ldots \qquad (10.2)$$

Looking at equation (10.1), we can conclude that the power of a lens is high when there is a large difference between the refractive indices (of lens versus medium) and/or its radii are small.

Taking all these into account we can calculate that **the cornea is the most refractive medium in the eye**, as:

- the difference between its index of refraction and that of air is higher than the difference between the index of refraction of the lens and the index of refraction of either of the two fluids that bathe the lens;
- the cornea is more curved than the lens (it has a lower radius).

Thus, the cornea has a refractive power of +40 to +43 D, while the lens is less refractive (with the refractive power of +17 to +20 D).

Together, both structures make up the overall refractive power of the eye of around +60 D. Although it has a lower power than the cornea, deviations from the normal functioning of the lens can lead to very unpleasant ocular manifestations. About half of the blindness cases in the world are due to **cataract**, a medical condition caused by a lack of transparency of the lens[5]. The symptoms include the visual sensation of diffused light, clouded vision or decreased contrast. However, there are surgical treatment methods that are successful in treating most forms of cataract (see the chapter on Physical factors in therapy).

Although the cornea does most of the refraction, the lens is responsible for the fine adjustments when focusing at various distances, as we will see later.

3.3. The model of the reduced eye

In order to simplify the laborious process of analyzing the refraction at each optical surface, a model called the *reduced eye* can be employed. A model that closely describes the optical behavior of the human eye was developed by Allvar Gullstrand, a discovery for which he was awarded the Nobel Prize in Physiology or Medicine in 1911.

Gullstrand's model was simplified by Emsley in 1952 (Figure 10.5) to consider all the refraction as occuring at a single refractive surface (the air-cornea interface) with its central point ~5.6 mm behind the cornea and ~17 mm in front of the retina (the nodal point, N in Figure 10.5). The reduced eye is considered homogenous, having the same refractive index as water (n = 1.333) and a total refractive power of ~ +60 D. It is useful in understanding the refractive errors of the eye (see later) and the calculation of the corrective lens power.

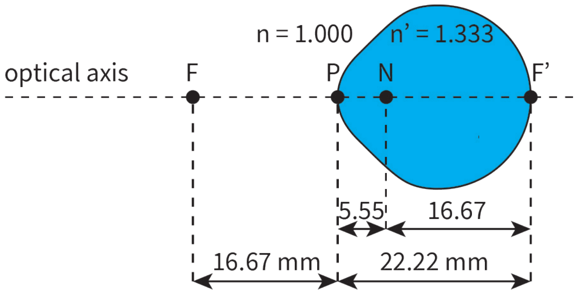


Figure 10.5. Gullstrand-Emsley model[6] of the reduced eye characterized by four cardinal points: F = first focal point of the eye, P = the anterior surface of the cornea, N = nodal point of the reduced eye, and F' = second focal point of the eye (on the fovea). The refractive index of the reduced eye (n') is simplified to 1.333 and, taking the refracting index of air as 1, the overall power of the eye is 60 D.

3.4. Adaptation (to various light intensities)

The amount of light that is focused on the retina is involuntarily controlled (pupillary light reflex). The iris behaves like a diaphragm, the pupil being its aperture with a diameter typically around 3 – 4 mm and varying between 1.5 and 8 mm in response to various light intensities:

- Light of high intensity triggers **pupillary constriction (miosis)** via contraction of the circular sphincter muscles which are controlled by the parasympathetic system (Figure 10.6, left panel);
- By contrast, low intensities of light trigger **pupillary dilation (mydriasis)** due to contraction of the radial dilator muscles, controlled by the sympathetic system (Figure 10.6, right panel).

The ability of the iris to adjust the diameter of the pupil is an important clinical tool for testing the integrity and normal functioning of the visual system. For example, the pupillary light reflex is commonly tested in emergency situations to assess the brainstem function.

3.5. Accommodation (to different distances)

As we have seen, in order to focus clear images of objects on the retina, the eye has to behave as

5 Actually, almost every person of more than 70 years old may experience some loss of the crystalline lens transparency (for example, due to denaturation of the transparent proteins).
6 Redrawn according to Emsley, H. H. (1953). *Visual Optics Volume I Optics of Vision* (Fifth ed.). London: Butterworths.

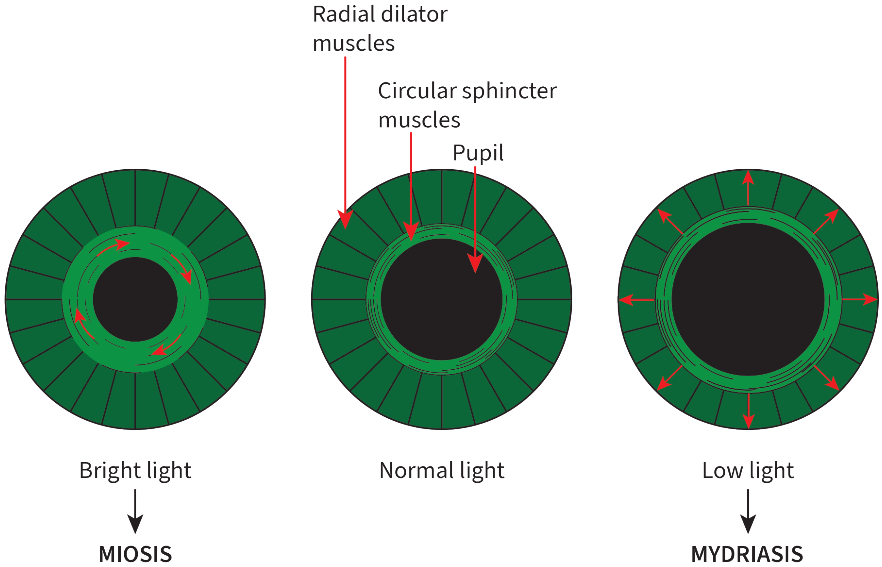


Figure 10.6. Adaptation of the pupil diameter to various light intensities through contraction of two types of iris muscles: circular sphincter muscles (their contraction leads to miosis) and radial dilator muscles (their contraction leads to mydriasis).

a spherical converging lens. In the Biophysics practical sessions and seminars, you have been introduced to the *thin lens equation*, which allows us to calculate the relation between the position of the object relative to the lens, the optical power of the lens, and the position of the formed image relative to the lens:

$$\frac{1}{x_2} - \frac{1}{x_1} = \frac{1}{f} = P \qquad (10.3)$$

where x_2 is the distance from the lens to the image (*image distance*), x_1 is the distance from the lens to the object (*object distance*), f is the focal distance and P is the optical power.

Note that the above equation is valid for the same sign convention mentioned for equation (10.1). In this particular case, where we have a real image formed through a converging lens, x_1 is negative, while x_2, f and P are positive.

For a regular lens, the optical power is a constant. As a consequence, if the object distance changes, the image distance will change accordingly. In other words, the image formed by an ordinary converging lens will be focused at a distance from the lens that depends on the position of the object relative to the lens. This is, obviously, unwanted in the eye, as clear vision can only be obtained for images that are focused exactly on the retina.

How does the eye, then, ensure that images of objects at different distances are all focused on the retina? The answer comes from the ability of the crystalline lens of the eye to change its shape, and consequently its power and focal distance according to the object distance, a process called **distance accommodation**.

There are several theories meant to explain this process among which the most widely accepted is the Helmholtz theory, which we will present in the following. Remember that the lens is held by the radially arranged suspensory ligaments (also called zonule of Zinn) which are connected to the ciliary muscle (via ciliary processes). The shape of the lens is the result of two opposing forces: the **elastic force**, which tends to maintain the lens spherical and the **tension** in the suspensory ligaments, which tends to flatten the lens. Because the ciliary muscle is like a ring around the lens, when it contracts, the insertion points of the suspensory ligaments on the ciliary body move toward the lens and the tension on the lens decreases, allowing it to become more convex (higher optical power) under the action of its elasticity. By contrast, when the ciliary muscle relaxes, the insertion points are pulled away from the lens, the tension on the lens increases and it becomes flatter (lower optical power).

As shown in Figure 10.7, a more convex lens enables the projection of images of nearby objects, as its optical power is higher, and thus its focal length is shorter. In other words, a more spherical lens bends the incoming rays of light more. On the other hand, when the lens of the eye is flattened, it is less convex, its optical power is lower, thus its focal length is longer. Otherwise put, a less spherical lens bends the incoming rays of light less to project on the retina images of distant objects.

The range of accommodation for the human eye has certain limits: a minimum and a maximum optical power (a minimum and a maximum focal length, respectively). These limits can be described also in terms of distances at which objects can be clearly seen:

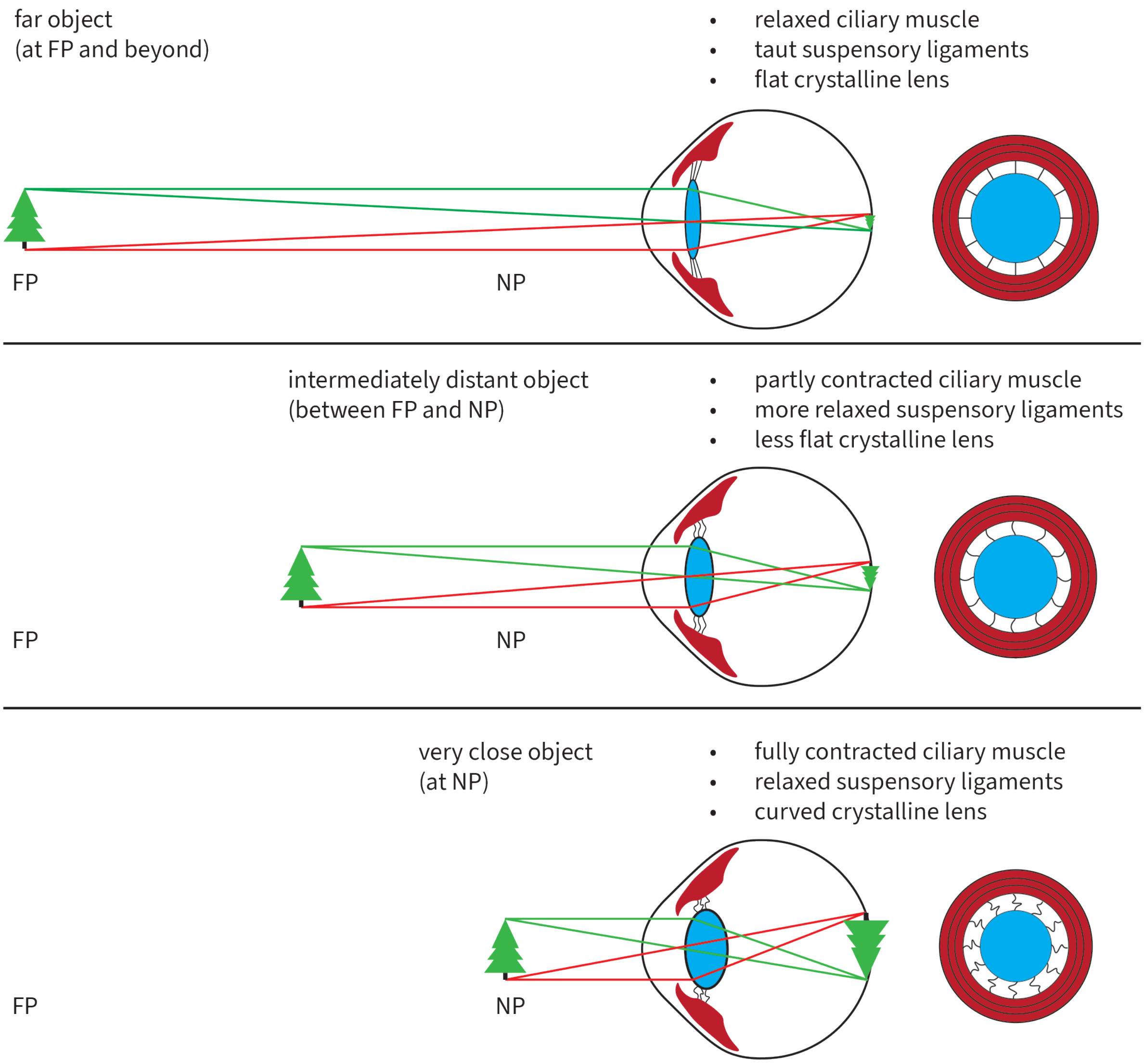


Figure 10.7. Accomodation for the emmetropic (normal) eye. Side and front view. In order for a clear image to be formed on the retina, the eye has to perform accomodation for objects situated closer than the far point (FP). This is performed by contracting the ciliary muscle, leading to a more curved crystalline lens. For objects closer than the near point (NP), a clear image cannot be formed on the retina. Images are not drawn to scale as the size of the eye is exaggerated relative to the distances to the FP and NP.

▸ The **far point** (*punctum remotum*) is defined as the longest distance between the objects and the eye at which their image is clear and projected on the retina. Ideally, the far point should be at infinity. In practice, a far point of **~6 m** is considered to be „normal", as there is no significant change in accommodation for distances higher than 6 m. The power of the eye is minimal, when viewing an object in the far point as described by a modified version of equation (10.3):

$$\frac{1}{far\ point} + \frac{1}{lens - retina\ distance} = P_{minimum} \quad (10.4)$$

▸ The **near point** (*punctum proximum*) is defined as the shortest distance between the objects and the eye at which their image is clear and projected on the retina. An average "normal" near point is considered to be **~25 cm**. The power of the eye is maximal when viewing an object placed at the near point, as described by a modified version of equation (10.3):

$$\frac{1}{near\ point} + \frac{1}{lens - retina\ distance} = P_{maximum} \quad (10.5)$$

Note that in equations (10.4) and (10.5), all distances should be taken as positive.

The word normal was written between quote marks, because these values are characteristic to what is considered a normal eye, which is more like an ideal eye, called **emmetropic eye**. However, many of us do not have these specific values for near and/or far points and we need to use eyeglasses or contact lenses, as will be described later, to help the eye to focus images of objects on the retina.

It should be pointed out that, although the ciliary muscle seems to be the active player in the accommodation process, the elasticity of the

crystalline lens is also important, especially for accommodation at short distances.

The difference between the highest and the lowest refractive power of the eye is the *amplitude of accommodation*. Its value may vary from one person to another and, for the same person is maximal during childhood and then decreases with increasing age (see the section on presbyopia).

4. REFRACTIVE ERRORS OF THE EYE

So far, we have only been concerned with normal vision (emmetropia). However, for some of us, the far point is closer than 6 m or the near point is at more than 25 cm. In addition to that, the eye may not behave like a spherical lens (for which the image of a point is also a point) but like a combination of a spherical and a cylindrical lens and, in this case, the image of a point may be a line. In all these examples images of various objects are not clear.

Although the human eye is far more complex than a simple lens, the previously described model of the reduced eye is very useful for explaining the geometry of normal vision (as presented in Figure 10.7) and also in understanding the physics of deviations from the characteristics of the emmetropic eye, commonly called *refractive errors*. We will discuss each type of refractive error in the following.

4.1. Myopia

Myopia (Figure 10.8) is a condition in which the images of distant objects appear blurred but images of close objects may be in focus. For this reason, this condition is also called *nearsightedness*. There are two major causes for myopia:

- a too long ocular axis (axial myopia);
- a too high refractive power of the eye (refractive myopia), which may be caused by an increase in the refractive index of the transparent media or, more commonly, by a an excessively curved cornea.

In either case, the image is not clear because the far point of the myopic eye is less than 6 m and, as equation (10.4) predicts, the power of the myopic eye is higher than the power of the emmetropic eye. Therefore, the images of objects which are further away are projected in front of the retina (Figure 10.8, top panel).

In order to correct myopia, a corrective lens must be used to decrease the overall power of eye: a **spherical diverging lens**. Using the corrective lens, the focal length will therefore be longer and the image will be projected on the retina (Figure 10.8, bottom panel).

4.2. Hyperopia (hypermetropia)

In **hyperopia** (Figure 10.9), the near point is not at 25 cm but at a longer distance. Thus, the images of close objects appear blurred because, as equation

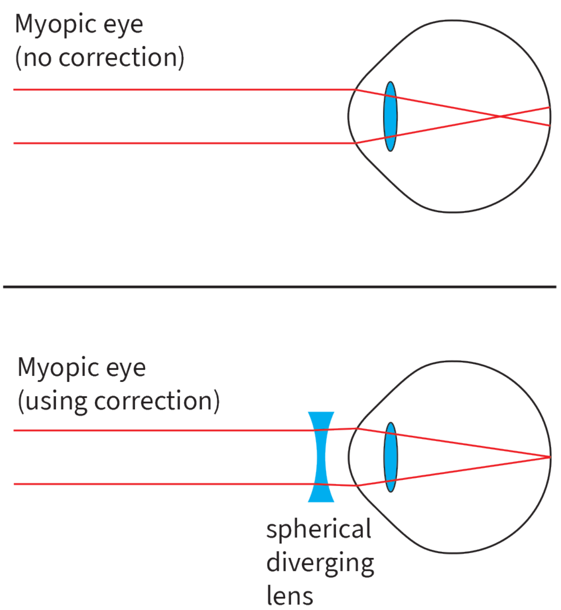


Figure 10.8. Myopia. Light rays coming from objects which are very far away are parallel when reaching the surface of the cornea. A myopic eye has a too high refractive power. Thus, it will focus those rays in front of the retina (top). In order to correct this defect, a spherical diverging lens is used which, together with the refractive media of the eye, will form an optical system that projects the image on the retina (bottom).

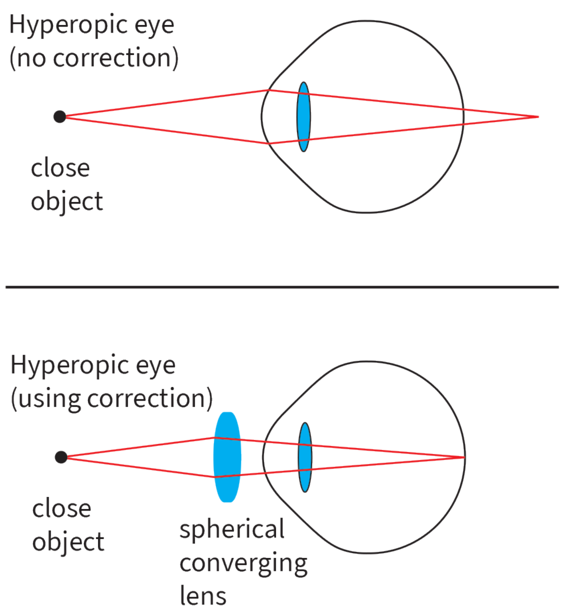


Figure 10.9. Hyperopia. The hyperopic eye has a too low refractive power. Thus, light rays coming from objects which are close to the eye are focused by the hyperopic eye behind the retina (top). In order to correct this defect, a spherical converging lens is used which, together with the refractive media of the eye, will form an optical system that projects the image on the retina (bottom).

(10.5) predicts, the power of the eye is lower than the power of the emmetropic eye, therefore the images of close objects are projected behind the retina (Figure 10.9, top panel). Images of far objects may be in focus and for this reason, this condition is sometimes called *farsightedness*. The two major causes for hyperopia are:

- a too short ocular axis (axial hyperopia);
- a too low refractive power of the eye (refractive hyperopia), which may be caused by a decrease in the refractive index of the transparent media or the cornea and/or the lens are not sufficiently curved.

Correction of hyperopia is achieved using a **spherical converging lens**. This increases the optical power of the eye, lowering the focal length and, thus, bringing the projection of the image on the retina (Figure 10.9, bottom panel).

4.3. Presbyopia

We previously presented (Figure 10.7) the process of accommodation and saw that that the process depends on the relative strength of the elasticity of the lens (which tends to increase the power) and the tension on the lens (which tends to decrease the power).

Measurements have shown that the largest amplitude of accommodation is achieved during childhood and it gradually decreases with age (Figure 10.10). This is probably due to fact that that the elasticity of the lens is highest at a small age. Thus, $P_{maximal}$ is at its highest value and, if we took another look at equation (10.5), we conclude that the corresponding near point at this age must be minimum. For example, an 8-year-old child has the amplitude of accommodation of 14 D corresponding to a near point of 7 cm. It slowly decreases with increasing age until around 40 years old (amplitude of accommodation is now around 4 D and the near point can still be 25 cm).

However, there is a steep decrease after that and usually the first sign is the need to move a book that you're reading further away in order to be able to clearly see the letters. At some point, your arm will be too short for you to move the book at an appropriate distance for reading, thus a corrective lens is needed.

This gradual loss of the ability to focus images of close objects is believed to be caused by a progressive decrease in the elasticity of the crystalline lens and is called **presbyopia**[8]. An eye with presbyopia is similar to the eye with hyperopia (Figure 10.9) and, likewise, it is corrected with **spherical converging lenses**.

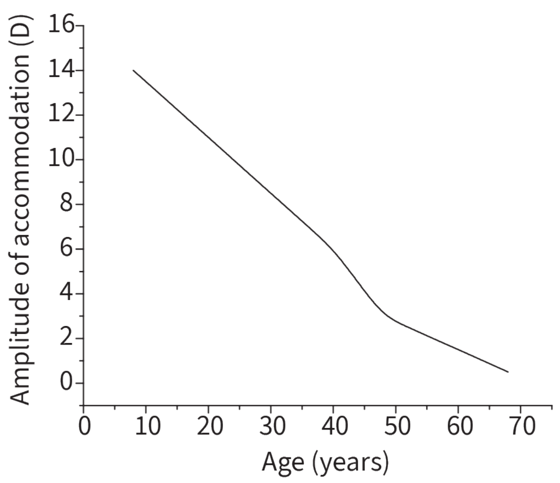


Figure 10.10. Decrease of the amplitude of accommodation with age.[7]

4.4. Astigmatism

In all the refractive errors discussed above, the eye behaved as a spherical converging lens but with an inappropriate optical power. There is another type of refractive error, **astigmatism** (Figure 10.11), caused by a non-constant curvature radius from one meridian to another (of the cornea, more often). The astigmatic eye behaves like a combination between a spherical and a cylindrical lens. Due to the cylindrical component

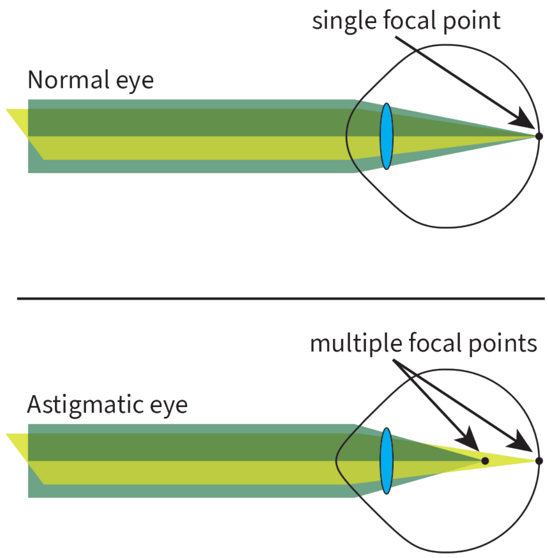


Figure 10.11. Astigmatism. In the normal eye (top), the rays of light originating from a single point of the object have a single focal point on the retina. In the astigmatic eye, these rays of light have multiple focal points, resulting in a blurry image.

7 Drawn using data from Brodie, S. E. (2020). *2020-2021 Basic and Clinical Science Course(tm) (BCSC), Section 03: Clinical Optics*: American Academy of Ophthalmology.

8 The word "presbyopia" is derived from a Greek word meaning "old eye".

of the lens, the image of a point is not always a point, but a line. Therefore, the rays of light that come from a single point in the object are not focused on a single point, but multiple focal points appear (Figure 10.11). This results in the image of the object appearing distorted (blurry).

In order to correct this defect, the cylindrical component of the eye must be canceled. To that end, **cylindrical lenses** having the same power (as the cylinder of the eye) but opposite sign are used.

Most often, because astigmatism is usually combined with myopia or hyperopia, the corrective lens must correct both defects, so it is a **spherocylindrical lens** with the appropriate power.

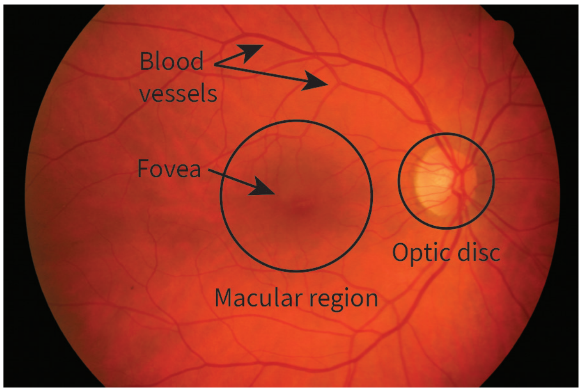


Figure 10.12. Image of the retinal surface (fundus of the eye) seen with an ophthalmoscope.[9] The main areas are: the macular region (macula lutea), with the fovea at its center and the optic disc.

5. VISUAL RECEPTION

It should be clear by now that, in order for us to see clear images, the light rays refracted by the optical system of the eye must be focused on the surface of the retina. In the following, the structures of the retina and their role will be described in order to understand how the light stimulus is converted into an electric signal (the phototransduction process). Finally, we will briefly describe color vision and deficiencies in color vision.

5.1. Structure of the retina

The retina is an extension of the central nervous system. It is composed of highly specialized neurons involved in the detection of light, the phototransduction process and in the analysis and processing of the neuronal signals.

Figure 10.12 shows the **macroscopic features** of the retina that can be observed using an ophthalmoscope:

- the **macula lutea** (or yellow spot), a small region toward the center of the retina relatively free of blood vessels;
- the **fovea**, the retinal area responsible for the sharpest vision, found in the middle of the macular region;
- the **optic disc**, the retinal area that transmits no visual information and is responsible for the presence of the **blind spot** in the monocular visual field.

If you are curious to find your blind spot, let's say for your left eye, try this: make a small X mark on the right side of a sheet of paper, then cover your right eye and look straight at the mark while slowly moving a sharp pencil toward it from your left side. At a certain position of the pencil, its tip is no longer visible (this is your blind spot). The light rays coming from it reach the optical disc, which has no photoreceptors. We are not normally aware of the blind spot because, in most of the cases, we have binocular vision. The absence of information in one eye is compensated by the other eye, because the visual fields overlap. Even if we had monocular vision, we wouldn't be aware of the blind spot, due to the brain's ability to perceptually fill in the gaps of information.

Figure 10.13 shows the **microscopic** structure of the retina. It contains:

- a **pigmented** part, the retinal pigmented epithelium;
- a **neuronal** part, consisting of five types of neurons: photoreceptors, horizontal cells, bipolar cells, amacrine cells and ganglion cells.

The **retinal pigmented epithelium** consists of a single layer of cells between the choroid and the photoreceptors. They have long processes extending into the photoreceptor layer so that the outer segment of each photoreceptor is surrounded by them. The activity of these cells is essential for the normal functioning of the photoreceptors; they *phagocytize the outer segments* of the photoreceptors and they *regenerate the photopigments* that were exposed to light. Also, the pigment melanin is present and it is important for absorbing photons that were not absorbed by the photopigments, thus reducing the unwanted reflections.

The **photoreceptors** are neurons specialized in the absorption of light (due to their content in photopigments) and in the conversion of the light signal into electric signal via the phototransduction process (which will be detailed later). Each photoreceptor is an elongated cell that has three

9 Modified from image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Budai, A., Bock, R., Maier, A., Hornegger, J., & Michelson, G. (2013). Robust Vessel Segmentation in Fundus Images. *International Journal of Biomedical Imaging*, 2013, 154860. doi:10.1155/2013/154860.

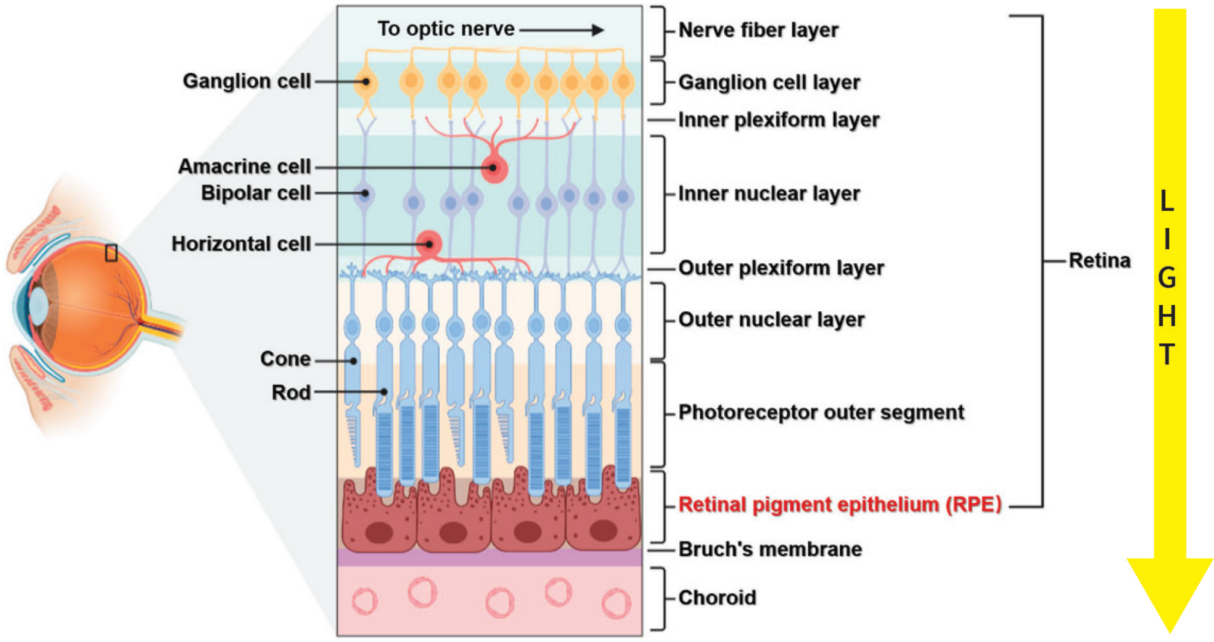


Figure 10.13. The microscopic structure of the retina.[10] Starting from the choroid toward the vitreous body, the main components of the retina are: the retinal pigmented epithelium, the photoreceptors (cones and rods), the horizontal cells, the bipolar cells, the amacrine cells and the ganglion cells. The outer segments of the photoreceptors contain the photopigments and, in order to reach them, light must cross all the other neuronal layers of cells, including the photoreceptors inner segments. The yellow arrow indicates the direction in which light enters and travels through the eye.

sections with distinct functions:

- **the outer segment** (containing the photopigments), positioned adjacent to the pigmented epithelium and connected via a ciliary stalk to the inner segment;
- **the inner segment** (containing the nucleus, mitochondria and other organelles), which continues with the synaptic terminal;
- **the synaptic terminal,** the communication site of the photoreceptors with the bipolar or the horizontal cells.

As shown in Figure 10.13, the outer segment is the last part of the photoreceptor reached by light and it is the site of the phototransduction process (see details below). Depending on the shape of their outer segments, the photoreceptor cells are of two types: **cones** and **rods**.

Each rod outer segment contains ~1000 membrane-delimited flat discs organized in a stack inside the cell (Figure 10.14, left panel).

Cone outer segments also have membranous discs; however, they are not delimited from the outer cell membrane but continuous with it (Figure 10.14, right panel). The disks contain photopigments, which are membrane bound proteins called opsins and a chromophore called retinal, a vitamin A derivative. Due to the arrangement of the membranous disks (parallel to surface of the retina) and to the high content in photopigments present in their membrane, the light absorption process is very effective.

The lifespan of the membranous disks is around two weeks. New disks are continuously formed at the basis of the outer segment, while older ones (near the pigmented epithelium) are phagocytized by the pigmented epithelial cells

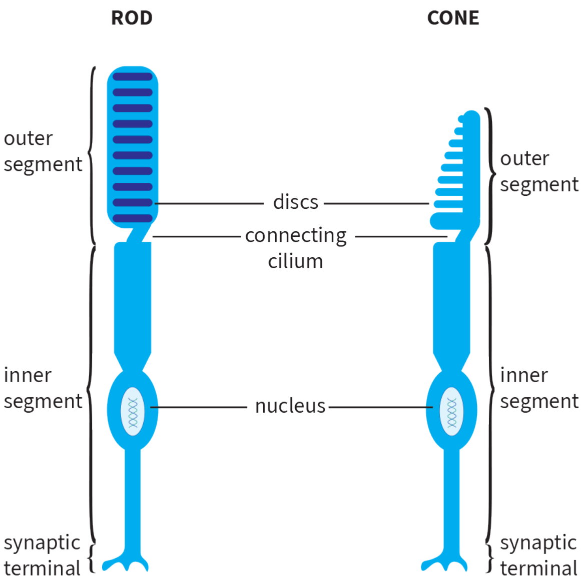


Figure 10.14. Types of photoreceptor cells. Left – rod; right – cone. The name of each type of cell originates from the shape of their outer segments.

10 Image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Yang, S., Zhou, J., & Li, D. (2021). Functions and Diseases of the Retinal Pigment Epithelium. *Frontiers in Pharmacology*, 12. doi:10.3389/fphar.2021.727870. Modified by adding arrow showing direction of light.

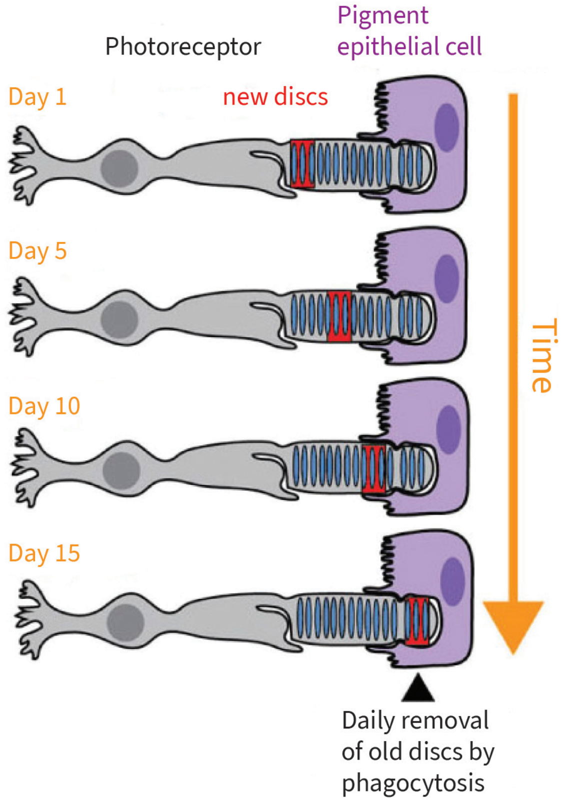


Figure 10.15. The lifespan of the membranous disks.[11] Newly formed disks are marked with red in Day 1. Over a period of about 15 days, they migrate toward the distal side of the outer segment, as newer disks are formed at its basis, while the older disks are phagocytized by the pigmented epithelial cells.

(Figure 10.15).

In case you were wondering why light must cross all those layers of cells in order to reach the photopigments in the outer segments of the photoreceptors, the most plausible answer is this: because it is crucial for the photoreceptors to have the outer segment close to the pigmented epithelium. In addition to the phagocytosis of the membranous discs, the pigmented epithelium is also important for the regeneration of the photopigment complexes after exposure to light. There is a continuous photopigment cycle between the outer segment of the photoreceptors and the pigmented epithelium. Also, the blood vessels within the choroid (remember that the pigmented epithelium is between the photoreceptor layer and the choroid) are the primary source of nourishment for the photoreceptors. An impaired relationship between the retinal pigmented epithelium and the photoreceptors may lead to serious vision problems. An example is *retinitis pigmentosa*, a group of hereditary disorders of the photoreceptors and the pigmented epithelium causing progressive vision loss.

Absorption of light by the photopigments from the outer segment triggers a cascade of biochemical events which leads to a gradual change in the photoreceptor's membrane potential, followed by a variation in the amount of neurotransmitter released at the synaptic termini. The most direct path for the neuronal signals to travel from the photoreceptors toward the optic nerve is a chain composed of three neurons (Figure 10.13): **photoreceptor → bipolar cell → ganglion cell**.

Ganglion cells are the only retinal cells with the ability of generating action potentials. Their axons form the **optic nerve**, which carries the visual information to the other structures of the central nervous system involved in vision. The other two types of retinal neurons modulate the transmission of these signals.

A small number of ganglion cells, called *intrinsically photosensitive retinal ganglion cells* contain the photopigment *melanopsin* and are sensitive to light. They play roles in regulating the circadian rythm and the pupillary light reflex.

Horizontal cells have processes that mediate a *lateral interaction* between the photoreceptors and the bipolar cells. The consequence of the horizontal cells' action is the sensation of contrast between areas with different light intensity.

The **amacrine cells** modulate the communication between the bipolar cells and the ganglion cells. There is a wide variety of amacrine cells, each type of cell bringing a distinct contribution to adjusting process of information which is passed to the ganglion cells and transmitted further on parallel pathways. Thus, the horizontal and amacrine cells can facilitate or inhibit (through lateral inhibition) the communication between the photoreceptors and the bipolar cells and between the bipolar cells and the ganglion cells, respectively. Their mechanism of action is complex and not fully understood. As an analogy, their effect is comparable to adjusting the contrast of your phone display.

5.2. The phototransduction process

Photoreceptor cells are a unique type of sensory cell. They do not depolarize but instead **hyperpolarize** when activated by their stimulus (light). Thus, **photoreceptor cells are neurons that do not generate action potentials, but work using local (graded) potentials**.

Let us explore how this process works.

In the absence of specific stimuli (light), the membrane of the photoreceptors is depolarized,

11 Modified from image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Kwon, W., & Freeman, S. A. (2020). Phagocytosis by the Retinal Pigment Epithelium: Recognition, Resolution, Recycling. *Frontiers in Immunology*, 11. doi:10.3389/fimmu.2020.604205.

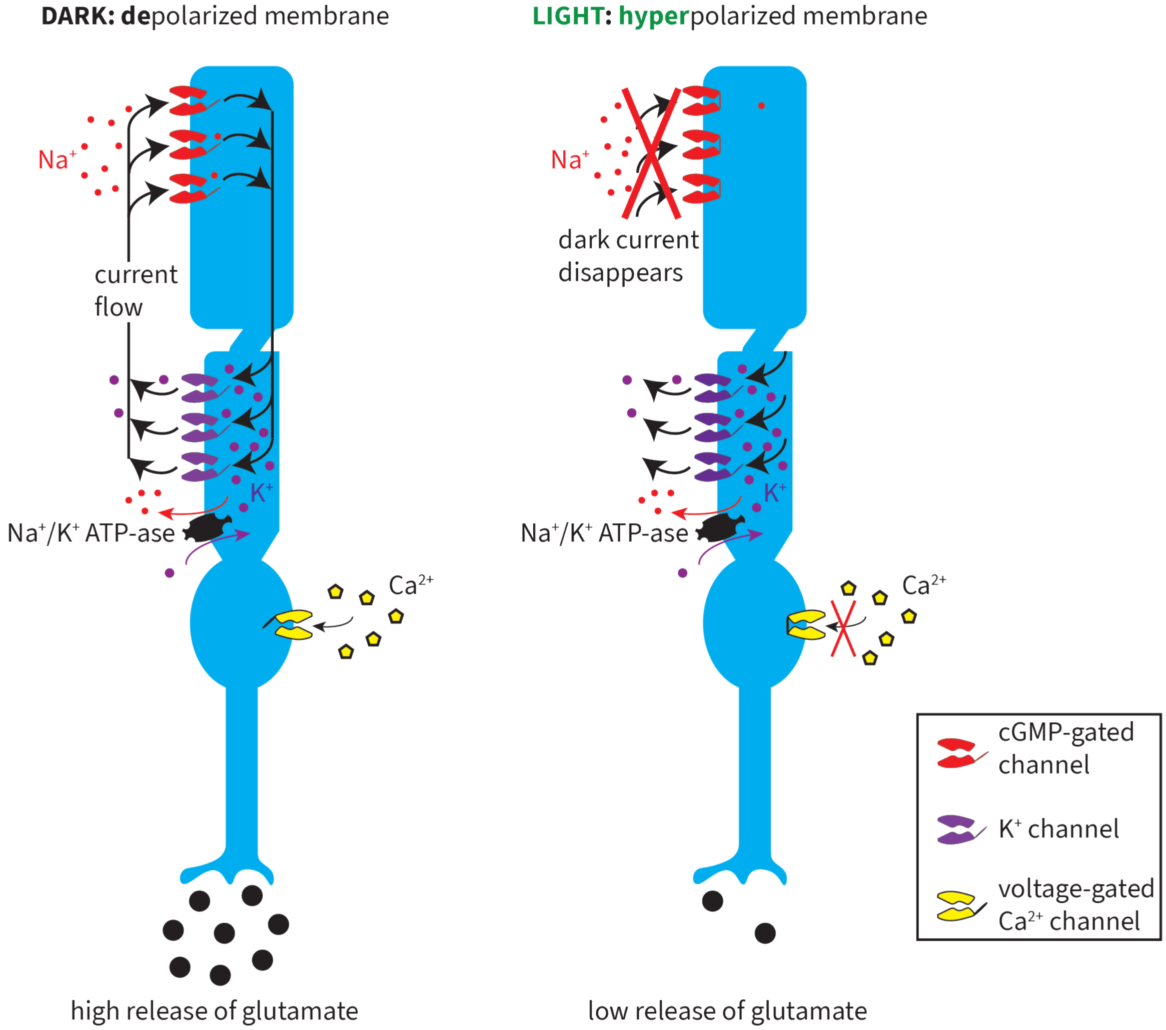


Figure 10.16. The ionic currents in a rod in the dark (left) or when exposed to light (right). In the dark, the membrane is depolarized due to a balance between the flow of positive charges out of the inner segment (K^+ ions) and the flow of positive charges into the outer segment, the dark current (mainly Na^+ ions). As long as the membrane stays depolarized, glutamate is released by the synaptic terminal at a high rate. When the rod is exposed to light, the dark current drops, while the K^+ current is insensitive to light. Thus, the membrane hyperpolarizes and the rate of glutamate release decreases accordingly. Note that, at all times, the Na^+/K^+ ATP-ase actively transports Na^+ and K^+ ions against their electrochemical gradients in order to maintain their intra-/extracellular concentrations constant.

with a membrane potential of approximately –40 mV. This depolarized state is a consequence of cationic membrane currents with opposite directions (Figure 10.16, left panel). On one hand, there is a K^+ current flowing out of the photoreceptor's inner segment through K^+ channels. On the other hand, there is the so called **dark current**, an inward current in the outer segment consisting mainly of Na^+ (but also Ca^{2+}) ions flowing through ion channels which are open only in the dark.

Therefore, in the dark (the resting state of the photoreceptors), there is a flow of positive charges going out of the inner segment which is opposed by a flow of positive charges going into the outer segment. The net result of these fluxes is a depolarized photoreceptor membrane. At this value of the membrane potential, voltage-gated calcium channels from the inner segment are open and the entry of Ca^{2+} ions stimulates the release of the neurotransmitter glutamate from the synaptic termini.

When activated by light, photoreceptors react by closing the cationic channels that mediate the dark current into the outer segment. The higher the intensity of light, the higher the number of channels that will close, thus the greater the reduction in the dark current.

In other words, under light exposure (Figure 10.16, right panel), the efflux of K^+ continues, while the influx of Na^+ and Ca^{2+} decreases. The result is a gradual hyperpolarization of the photoreceptor's membrane down to about –65 mV, followed by a reduction in the number of voltage-gated Ca^{2+} channels which are open and a corresponding decrease in the amount of glutamate that is released.

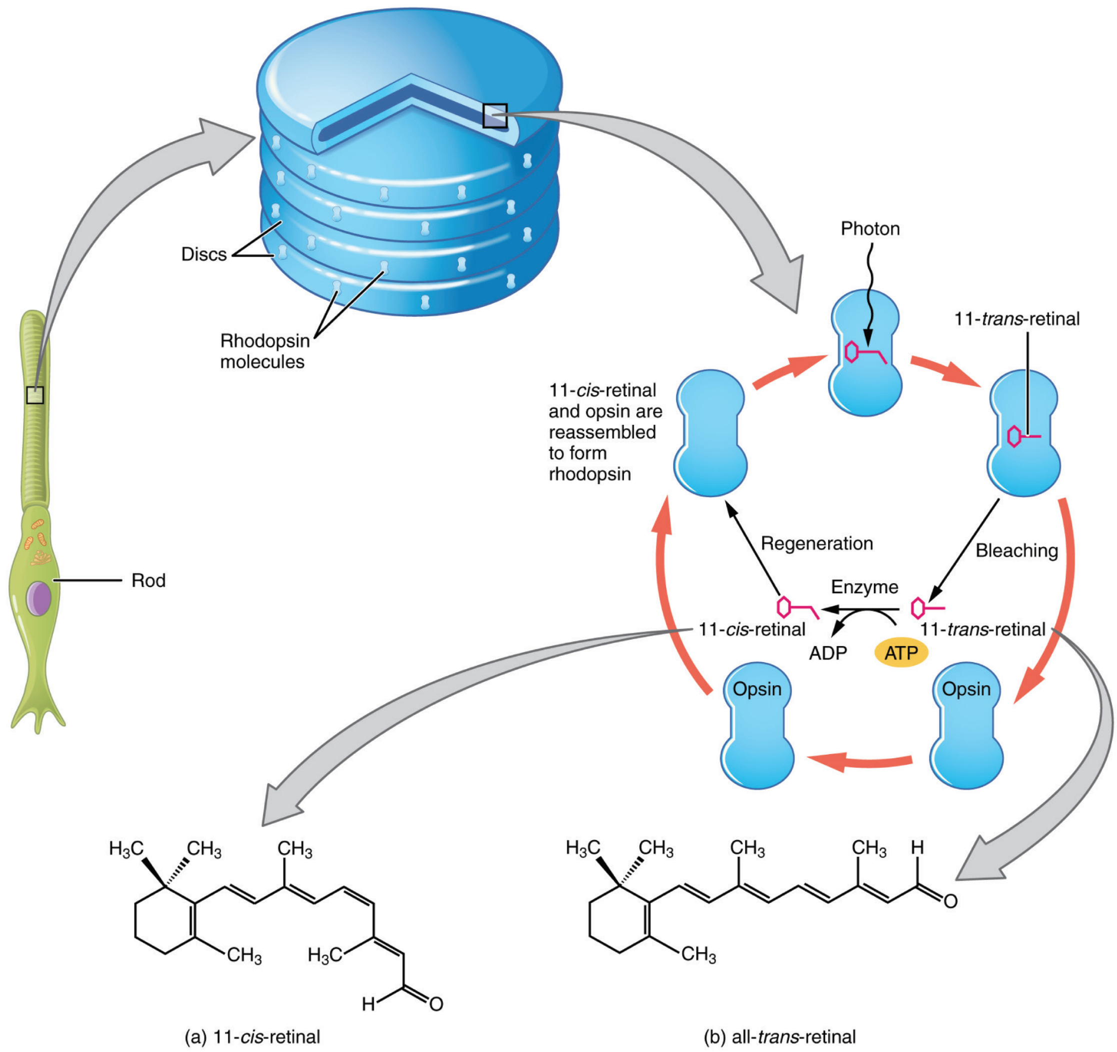


Figure 10.17. The retinoid cycle in a rod.[12] Absorption of a photon by a rhodopsin molecule triggers a photoisomerization process in which 11-*cis* retinal is converted into all-*trans* retinal, followed by all-*trans* retinal dissociation from opsin, a process called bleaching (it cannot absorb light in this state). All-*trans* retinal is transported from the cytosol of the outer segment into the pigmented epithelium, where it is enzymatically reconverted into the 11-*cis* conformation and then transported back into the rod's outer segment where it recombines with an opsin molecule to form the light sensitive rhodopsin. There are several intermediate forms of retinal between all-*trans* retinal and 11-*cis* retinal in the regeneration phase of the retinoid cycle, but these details are beyond the scope of this chapter.

But how is light stopping the dark current? We need to take a closer look at the molecular makeup of the outer segment. In the following we will talk mainly about rods, because most of the information we have about the molecular mechanism of phototransduction was obtained from experiments on rods. We need to go back to the membranous discs from the photoreceptor's outer segment, which contain a high number of photopigments. Also remember that this outer segment is adjacent to the pigmented epithelium.

Rods contain **rhodopsin**, a pigment composed of a transmembrane protein called **opsin** forming a pocket inside which there is a **retinal** molecule (the chromophore). In the dark, one of the double bonds of the retinal molecule is in the *cis* conformation (11-*cis* retinal). When exposed to light, 11-*cis* retinal can absorb a photon, changing its conformation into all-*trans* retinal (Figure 10.17). There are two immediate consequences of this photoisomerization process:

- changes in the opsin conformation which initiate a cascade of biochemical events;
- dissociation of the all-*trans* retinal from the opsin and its diffusion into the cytosol.

When it is in all-*trans* configuration, retinal can

12 Image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). *Anatomy and Physiology 2e*. Retrieved from *https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction*

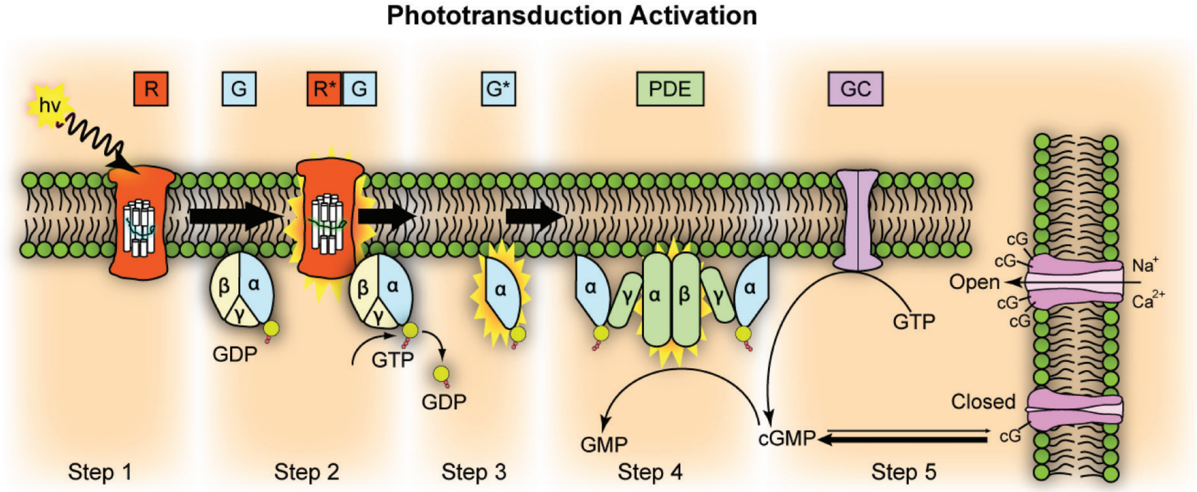


Figure 10.18. The main steps of the phototransduction process.[13] **Step 1**: A molecule of rhodopsin (R) absorbs a photon (hν) and is activated (R*). **Step 2**: R* interacts with transducin (a G protein), stimulating the exchange of GDP for GTP in its alpha (α) subunit. **Step 3**: The α subunit of transducin with GTP attached (G*) dissociates from the complex and diffuses along the disc membrane. **Step 4**: G* activates phosphodiesterase (PDE) which catalyzes cGMP hydrolysis to GMP. **Step 5**: The decrease in cGMP concentration causes the cGMP-gated channels to close.

no longer contribute to the absorption of light and it must be converted back to the 11-*cis* isomer in order to be able to do that again. The process is not spontaneous and it requires energy consumption (e.g. ATP hydrolysis through enzymatic activity). The enzymes specialized in reconversion to 11-*cis* retinal are present in the pigmented epithelium. Thus, the all-*trans* isomer is transported by specialized proteins into the pigmented epithelium and then, the newly formed 11-*cis* retinal is transported back into the outer segment where it recombines with opsin in the membrane of the disks and it is ready for the absorption of a new photon. A brief description of this cycle of events (the retinoid cycle) is depicted in Figure 10.17. The **retinoid cycle** has a crucial importance for the maintenance of light sensitivity of photoreceptors, especially in bright light conditions; due to rapid regeneration of the pigment, there are always some photopigment molecules available for light absorption.

We still haven't answered to the question about how light is stopping the dark current. In order to do that, we need to describe the actual **phototransduction** process (Figure 10.18). Remember that photoisomerization of retinal induces a change in the opsin configuration. This change activates a series of events that take place within the disc membrane. First, active rhodopsin stimulates a G protein (called transducin), a peripheral membrane protein, which in turn activates another peripheral membrane protein. The latter is an enzyme called phosphodiesterase which catalyzes the hydrolysis of cyclic guanosine monophosphate (cGMP), thus reducing the concentration of cGMP.

The ion channels from the membrane of the outer segment which mediate the dark current (Figure 10.16) are a type of ligand gated channels called cyclic nucleotide-gated channels. In the dark, they are open due to a high intracellular concentration of cGMP which binds to specific sites on the channels, activating them.

So, what happens if cGMP is hydrolyzed by phosphodiesterase? First, its concentration starts decreasing, reducing the number of cGMP molecules available to bind cGMP-gated channels. Then, cGMP molecules dissociate from the cGMP-gated channels. In the absence of their ligands, the ion channels close, thus stopping the dark current. Thus, cGMP is the intracellular messenger that mediates the communication between the disc membrane and the membrane of the outer segment.

The biochemical chain of events initiated by the absorption of a photon is characterized by a huge signal amplification in rods. To have an idea: a single activated rhodopsin molecule activates ~800 transducin molecules (about 8% from the total tranducins in the disc membrane), each transducin activates one phosphodiesterase, however each enzyme can hydrolyze 6 cGMP molecules, leading to the closure of about 200 cGMP gated ion channels (about 2% of the total channels open in the dark), which causes a change of about 1 mV in the membrane potential. However, the

13 Image available under a GNU Free Documentation License (*https://commons.wikimedia.org/wiki/Commons:GNU_Free_Documentation_License,_version_1.2*) from Jason J. Corneveaux (*https://en.wikipedia.org/wiki/File:Phototransduction.png*).

duration of this amplification process is limited by other biochemical processes, in order to allow rhodopsin to return to its inactive state (through the retinoid cycle described above).

5.3. Function and distribution of rods and cones

The **sensitivity to light**, **spatial resolution** and **detection of colors** are three important aspects of vision. Rods and cones bring different contributions. Due to their ability to amplify the signal, rods are extremely sensitive to light, however they do not contribute much to the spatial resolution and not at all in color sensing. In other words, rods do not help us in seeing the details or the colors of objects around us. The cones, however, enable us to see colors and details, but they have low sensitivity to light (a cone needs to absorb about 100 photons in order to produce a response comparable to the response of a rod following the absorption of only one photon). Depending on the intensity of illumination, three types of vision can be distinguished:

- **Scotopic** vision (poor resolution, no color), mediated only by **rods**, when the intensity of light is very low (for example, a cloudy night, with no artificial source of light on);
- **Mesopic** vision, mediated by both rods and cones, when the intensity of light is slightly higher (for example, at twilight);
- **Photopic** vision (detailed, colored vision), mediated only by **cones**, at normal indoor light or sunlight.

We can conclude that the contribution of cones is dominant in most situations in which we are looking at something. For example, in daylight, rods do not contribute to vision because they are saturated; all their nucleotide gated channels are closed, so they can no longer change their membrane potential in response to changes in illumination.

Although what we normally see in daylight or at normal indoor lighting is based on the activity of cones, there are much more rods in the retina (about 90 million rods versus 4.5 million cones). In other words, there are 20 times more rods than cones in the retina. As a consequence of that, the density of cones is lower in most of the retinal areas, except for the fovea, where it increases steeply and reaches a maximum in its center (Figure 10.19).

Cones and rods also differ in their synaptic connections. For example, the pathway from cones to ganglion cells involves a certain type of bipolar cells, while the rods signal to another type of bipolar cells, although in the end the signals can converge onto the same ganglion cell. The degree of convergence is another important

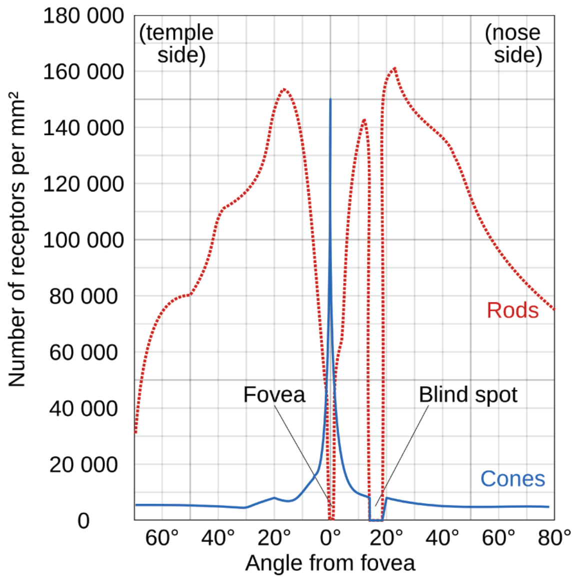


Figure 10.19. The density of cones and rods represented as the number of photoreceptors per mm^2 versus angle from the fovea, for a section through a human left eye.[14] The center of the fovea is taken as a reference and considered to be at an angle of 0°. In this spot, there is a maximum density of cones (solid blue line) which decreases steeply with increasing angle. Rods dominate most of the areas, with a sharp decline in the fovea until they reach 0 density in the center of it (red dotted line). Both cones and rods are missing from the optic disk. This absence causes the perception of the blind spot.

difference between cones and rods. For example, each ganglion cell from the center of the fovea receives input from one bipolar cell, which, in turn, receives input from only one cone. This ratio of 1:1:1 enables the high acuity vision. Rods, however have a much higher degree of convergence; many rods signal to a rod bipolar cell and many rod bipolar cells contact a certain amacrine cell which mediates communication with a certain ganglion cell. So, it is the higher convergence which make rods better at detecting light but worse at spatial resolution.

5.4. Information processing by retinal circuitry

There are many more photoreceptors than fibers in the optic nerve, therefore the image focused on the retina cannot be transmitted point by point to the brain. Instead, the visual signal is processed from these early stages, before action potentials are generated into the ganglion cells and transmitted toward the brain. It has been found that the retinal cells "compress" the information received from the sensory cells (conceptually like

14 Image available under a Creative Commons License (*https://creativecommons.org/licenses/by-sa/3.0/deed.en*) from Cmglee (*https://commons.wikimedia.org/w/index.php?lang=und&title=-File%3AHuman_photoreceptor_distribution.svg*).

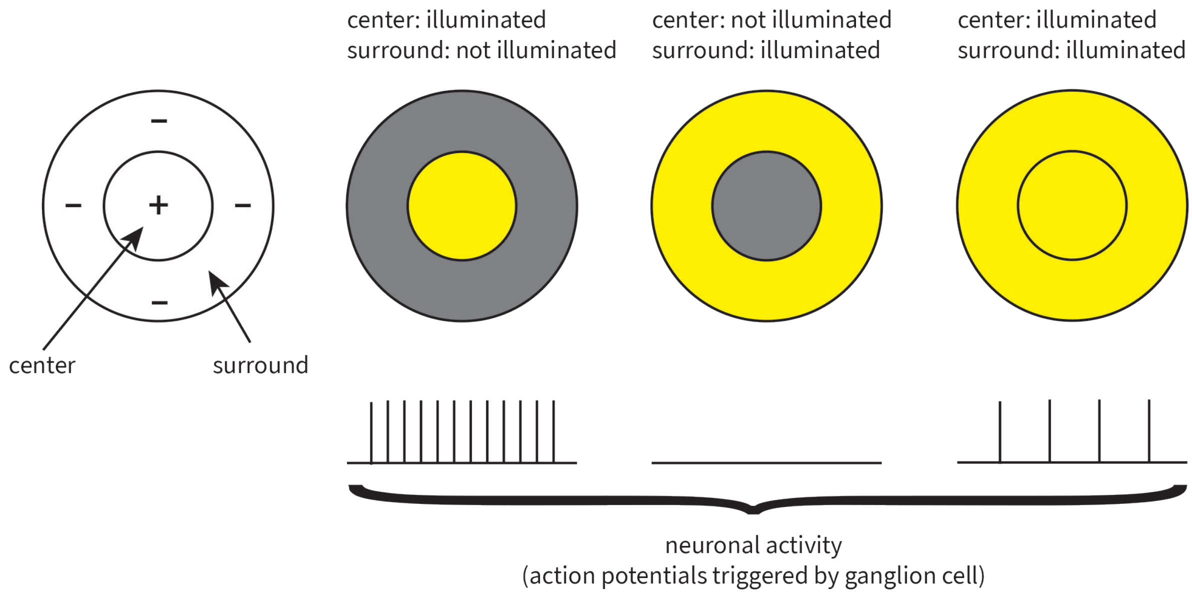


Figure 10.20. Example of ganglion cell ON-center/OFF-surround. The circular receptive fields with the two different areas (center and surround) are shown in the leftmost panel. The next three panels show the effects of illuminating different regions of the receptive field on the firing of action potentials by the ganglion cells. If the center of the receptive field is illuminated, while the surround is not, action potentials will be triggered with a high frequency. If only the surround is illuminated, while the center is not, no activation of the ganglion cell will occur. Finally, if both the center and the surround are illuminated, action potentials will be triggered, but at a lower frequency than in the first case.

compressing a computer file); the actual processing is quite complicated, but the general idea is that different features of the input (contrast, size, motion, color) are processed and sent separately to the brain via different neuronal pathways).

Experiments have shown that a bipolar or a ganglion cell responds to light shining on a specific retinal area, which is circular and is called receptive field. The size of the receptive filed varies with distance from the fovea: the peripheral receptive fields are larger than those in the fovea. Thus, the degree of convergence is correlated with size of the receptive field.

The response of a ganglion cell depends on the location of the stimulus within the receptive field. One of the models used to explain this behavior is called the "center – surround model". According to this model, each receptive field has a center that responds differently than the surrounding area, thus there can be two types of ganglion cells: ON-center/OFF-surround or OFF-center/ON-surround. For teaching purposes, Figure 10.20 presents an oversimplified situation of an ON-center/OFF-surround ganglion cell. The activity of bipolar and ganglion cells is much more complex and it depends on their synaptic connections and on the type of glutamate receptors present on their membrane, but these mechanisms[15] are beyond the scope of this book. The consequence of this organization is an increase in contrast between the illuminated area and its surroundings, thus increasing visual acuity.

5.5. Cones and trichromatic vision

Cones and rods also differ in the type of photopigment they contain. Rods contain a single type of pigment (**rhodopsin**), but the cones have three, called **photopsins** (when not bound to retinal) or **iodopsins** (when bound to retinal). As consequence of that, there are three types of cones, depending on the pigment they contain:

- **S** cones or **blue** cones – contain the pigment **cyanolabe** that absorbs light of *short* wavelengths);
- **M** cones or **green** cones – contain the pigment **chlorolabe** which absorbs light of *medium* wavelengths;
- **L** cones or **red** cones – contain the pigment **erythrolabe** that absorbs light of *long* wavelengths.

All four photopigments consist of the membrane bound opsins and the light absorbing retinal. Each type of photopigment contains a different opsin. Consequently, each type of opsin binds to the retinal in a specific way, which gives the photopigment a specific absorption spectrum.

15 If you are curious to know more, detailed description of these mechanisms is presented in chapter 12 of the following book: Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A. S., R.D., M., . . . White, L. E. (2018). *Neuroscience, 6th Edition*. New York: Sinauer Associates.

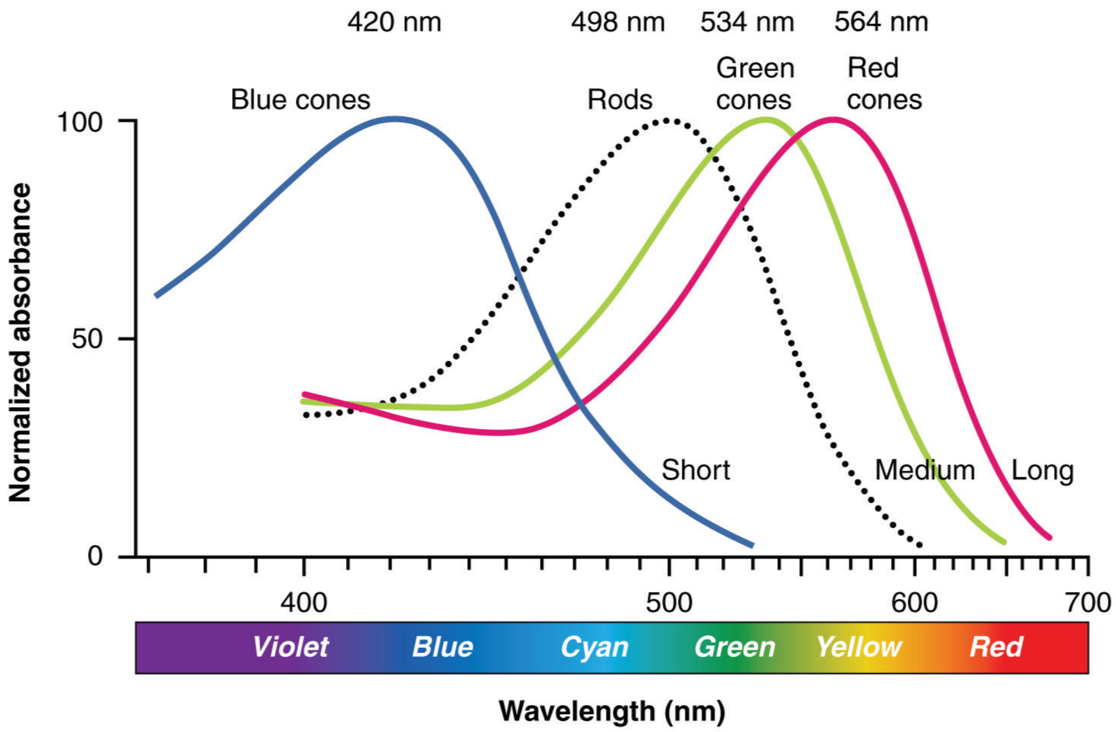


Figure 10.21. Absorption spectra of photopigments present in S cones (blue), M cones (green), L cones (red) and rods (black, dotted line).[16]

Figure 10.21 shows the absorption spectra for the all four types of photopigments.

The three types of cones are in different proportions (for example, S cones are only 5 – 10 % of the total number of cones), which may vary from one individual to another and have a mosaic distribution across the retina.

The blue/green/red name implies the color perceptually associated with the wavelength of light that best excites the cones (maximal absorbance in Figure 10.21). It does not mean, however, that they transmit color information related to that wavelength alone. The perception of color is a complex mechanism that involves both peripheral and central processing of the visual information. For example, an important aspect is the relative activity of the three types of cones, because of which humans have **trichromatic vision**.

When light of all wavelengths enters the eye, all the three types of cones are equally stimulated and trigger a cascade of events culminating with the perception of white color. If only some wavelengths are present, cones are stimulated at different degrees, and this is interpreted as a certain color.

For example, using Figure 10.21, we can see that, if the eye receives a ray of monochromatic light at 540 nm, this will not stimulate the S cones at all, it will maximally stimulate the M cones and it will moderately stimulate the L cones. Our brain will interpret the color of this light as green.

If, instead, a monochromatic ray of light at 620 nm is received by the eye, this will not stimulate the S cones, it will only weakly stimulate the M cones and it will moderately stimulate the L cones. The color interpreted by the brain will be red for this wavelength.

5.6. Dyschromatopsia

There are genetic disorders in which one or more types of cones are lacking or nonfunctional and lead to a defective color vision called dyschromatopsia or color blindness.

The inability to see any color (total color blindness = *achromatopsia*) is very rare. Most commonly, color deficiency occurs when L or M cones are missing – those affected have only two types of functional cones, thus they are dichromats. The most common form of color blindness is a condition called *deuteranopia*, caused by the absence of M cones. A less common condition is *protanopia* (due to the absence of L cones). *Tritanopia* is the least common (absence of S cones).

People missing only M or L cones have difficulty in distinguishing green from red, as shown

16 Image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). *Anatomy and Physiology 2e*. Retrieved from *https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction*



Figure 10.22. Comparison between an image viewed using normal, thrichromatic vision (left) and that viewed by a person suffering from deuteranotpia (red-green color blindness). Notice that rightmost toy mouse becomes almost indistinguishable in color compared to the other mouse in the right image. Also, the color of the pillow seems the same as that of the left toy mouse in the right panel. Picture shows O.C.'s cat (Salsa).

in Figure 10.22). Because the genes coding for the opsins specific to these cones are on the X chromosome, and since men have only one X chromosome, red-green color blindness (daltonism) is much more common in men than in women.

REFERENCES

Angée, C., Nedelec, B., Erjavec, E., Rozet, J.-M., & Fares Taie, L. (2021). Congenital Microcoria: Clinical Features and Molecular Genetics. *Genes*, 12(5), 624. Retrieved from https://www.mdpi.com/2073-4425/12/5/624

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2012). *Biochemistry. Seventh Edition*. New York: Freeman and Company.

Boron, W. F., & Boulpaep, E. L. (2017). *Medical Physiology* (3 ed.). Philadelphia: Elsevier.

Brodie, S. E. (2020). *2020-2021 Basic and Clinical Science Course(tm) (BCSC), Section 03: Clinical Optics*: American Academy of Ophthalmology.

Budai, A., Bock, R., Maier, A., Hornegger, J., & Michelson, G. (2013). Robust Vessel Segmentation in Fundus Images. *International Journal of Biomedical Imaging*, 2013, 154860. doi:10.1155/2013/154860

Ecker, J. L., Dumitrescu, O. N., Wong, K. Y., Alam, N. M., Chen, S. K., LeGates, T., . . . Hattar, S. (2010). Melanopsin-expressing retinal ganglion-cell photoreceptors: cellular diversity and role in pattern vision. *Neuron*, 67(1), 49-60. doi:10.1016/j.neuron.2010.05.023

Emsley, H. H. (1953). *Visual Optics Volume I Optics of Vision* (Fifth ed.). London: Butterworths.

Faralli, J. A., Filla, M. S., & Peters, D. M. (2019). Role of Fibronectin in Primary Open Angle Glaucoma. *Cells*, 8(12), 1518. Retrieved from https://www.mdpi.com/2073-4409/8/12/1518

Franklin, K., Muir, P., Scott, T., & Yates, P. (2019). *Introduction to Biological Physics for the Health and Life Science*s: Wiley.

Giancoli, C. G. (2008). Physics for Scientists & Engineers with Modern Physics, 4th Edition: Pearson Education.

Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). *Anatomy and Physiology 2e*. Retrieved from https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction

Guyton, A. C., & Hall, J. E. (2005). *Textbook of Medical Physiology. Eleventh Edition*. Philadelphia: Elsevier.

Kwon, W., & Freeman, S. A. (2020). Phagocytosis by the Retinal Pigment Epithelium: Recognition, Resolution, Recycling. *Frontiers in Immunology*, 11. doi:10.3389/fimmu.2020.604205

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A. S., R.D., M., . . . White, L. E. (2018). *Neuroscience, 6th Edition*. New York: Sinauer Associates.

Uddin, N., & Rutar, M. (2022). Ocular Lymphatic and Glymphatic Systems: Implications for Retinal Health and Disease. *International Journal of Molecular Sciences*, 23(17), 10139. Retrieved from https://www.mdpi.com/1422-0067/23/17/10139

Yang, S., Zhou, J., & Li, D. (2021). Functions and Diseases of the Retinal Pigment Epithelium. *Frontiers in Pharmacology*, 12. doi:10.3389/fphar.2021.727870

Zong, Y., Gao, Q. Y., & Hui, Y. N. (2022). Vitreous function and intervention of it with vitrectomy and other modalities. *Int J Ophthalmol*, 15(6), 857-867. doi:10.18240/ijo.2022.06.02

CHAPTER 11

BIOPHYSICS OF HEARING

Prerequisite knowledge

- Waves and their characteristics
- Action potentials
- Pressure and its units of measurement

Our sense of hearing depends on the physics of the sound waves, but also on the physiology of our ears and the psychology of our brain. In general, the word "sound" is used to name both the stimulus and the sensation evoked by the detection of the stimulus. However, the two terms are different from a scientific point of view. The sound stimulus is a mechanical wave, with properties that do not depend on who is listening (an objective phenomenon), whereas the sensation of sound is a subjective phenomenon. In order to make the distinction between the two, in this chapter, we are going to refer to the stimulus as "**acoustic signal**"[1] and we will use "**sound**" when referring to the sensation.

1. THE ACOUSTIC SIGNAL

1.1. Mechanical longitudinal waves

The acoustic signal is a physical phenomenon that consists of alternating compressions and rarefactions which propagate through an elastic medium (e.g. air, water) as a *mechanical longitudinal wave*. If this sounds complicated, let us define each term from the above statement in order to understand its meaning.

First, a **mechanical wave** is a disturbance from an equilibrium state that travels through a certain medium, via oscillation of the particles composing the respective medium. Imagine a metallic spring or a slinky toy with one end attached to a wall and the other end into your hand. You keep your hand still, so the spring is in a horizontal position and not moving. We will consider this the equilibrium position of the spring. Now, imagine that you move your hand up and down, the movement of your hand would determine the particles of the spring in contact with your hand to also move up and down and this oscillation will be transmitted to the next particles in the spring which will also oscillate up and down and thus the disturbance you have created propagates from your hand toward the wall (Figure 11.1, top panel). The direction of oscillation (vertical) and the direction in which the wave is propagating (horizontal) are perpendicular and the wave is called a **transverse wave**.

Next, you go back to the equilibrium position and then you move your hand horizontally by pushing the spring and withdrawing your hand.

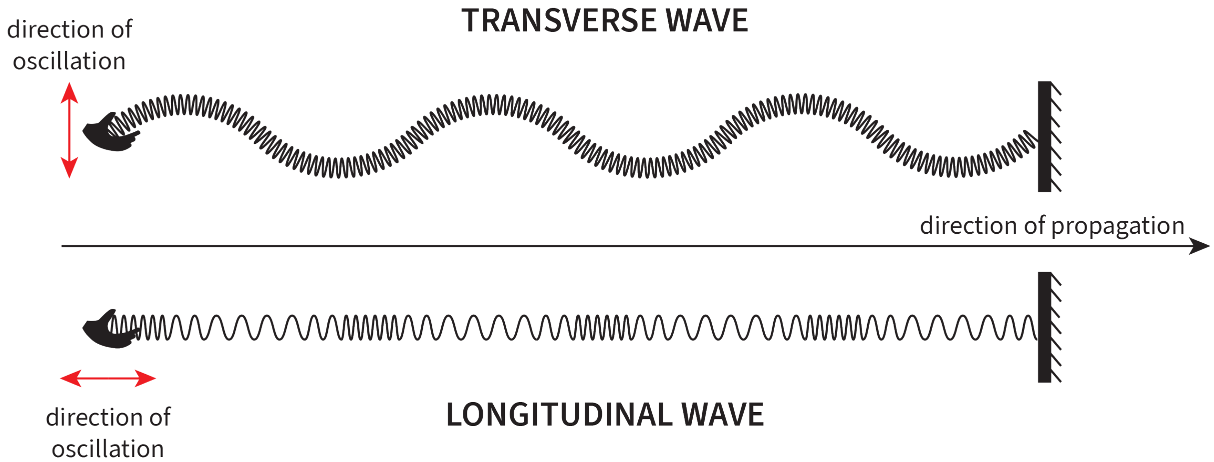


Figure 11.1. Mechanical waves produced in a spring (slinky toy) fixed at one end. A transverse wave (top panel) and a longitudinal wave (bottom panel) can be produced in a spring by moving the spring differently.

1 *Acoustics* is a branch of physics that studies mechanical waves. The term is derived from the Greek "akoustos", meaning "heard".

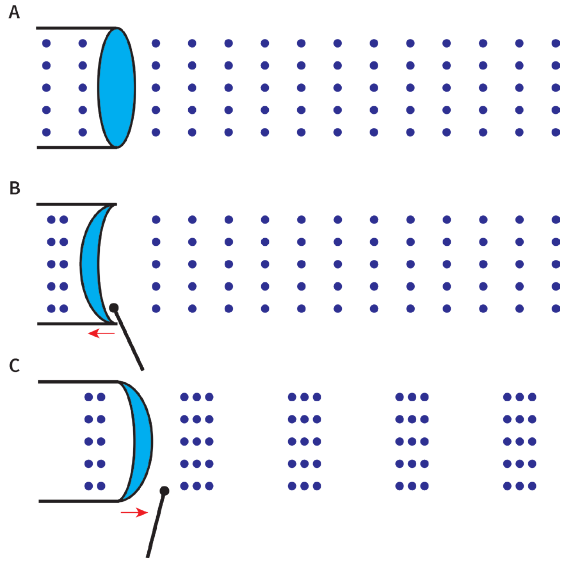


Figure 11.2. A drum produces an acoustic (longitudinal) wave. Molecules of gas in the air are represented as circles. A, the drum at equilibrium.[2] B, the drumhead is hit by the drumstick, compressing the air behind it. C, the drumhead recoils, producing an acoustic wave that propagates outward.

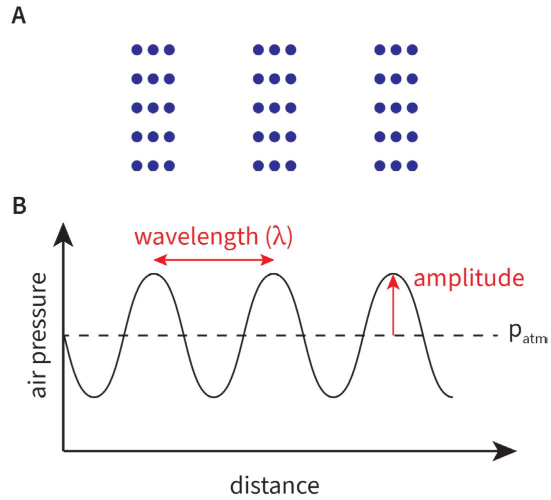


Figure 11.3. Longitudinal wave in the air (A) and its graphical representation of the wave at certain moment in time as deviation from the ambient (equilibrium) atmospheric pressure (B). Note that the differences in distribution of the air molecules in panel A are exaggerated for the purpose of the picture, as the actual deviations in pressure from p_{atm} are normally very small (fractions of a percent).

This movement will cause a horizontal oscillation of the spring particles which will be transmitted towards the wall (also horizontally); the oscillations are now parallel to the direction of the wave and this is an example of a **longitudinal wave**. If you took a snapshot of the oscillating spring, the picture would show areas where the coils are close together (*compressions*) alternating with regions where the coils are far apart (*rarefactions* or *expansions*) (Figure 11.1, bottom panel). Therefore, you can say that the longitudinal wave is transmitted by alternating compressions and rarefactions that propagate along the spring.

Before moving on, let's consider some of the main differences between mechanical waves and the electromagnetic waves that we studied in previous chapters:

- **Electromagnetic waves do not require a medium in order to propagate (they can propagate through vacuum), while mechanical waves require a medium for propagation**. Thus, mechanical waves cannot propagate through vacuum;
- Furthermore, while electromagnetic waves (light) are always transverse waves, mechanical waves can be both transverse or longitudinal (Figure 11.1).

The acoustic signals that we can detect as sounds are mechanical longitudinal waves. They are generated by vibrating objects (a musical instrument, the vocal cords of a person speaking, etc.) and propagate through air. The position of air molecules before the vibration of the object is taken as reference (similar to the equilibrium position of the spring in the example given above).

For example, when a drummer hits the head of a drum, it vibrates, alternatively compressing and rarefying the air just in front and just behind it (Figure 11.2). The molecules of air will oscillate in the same direction as the drum head, their oscillation will be transmitted to their neighboring molecules (i.e. through collisions), and thus producing a longitudinal wave that travels outward in the air until it reaches your ears. This is an example of an acoustic signal. Note that **each particle of the medium in which a longitudinal wave propagates oscillates over a very small distance, whereas the wave itself can travel long distances**.

As Figure 11.2 shows, in the compression areas, the density (and therefore the pressure) of air molecules is higher than in the rarefaction areas. In general, acoustic waves can be described in two ways: in terms of displacement of the molecules at each position in the medium (see below) or in terms of pressure at each position in the medium. If we plot the pressure of air molecules versus distance from the vibrating object at a given instant, we obtain a mathematical representation of the longitudinal wave, a sinusoidal wave for which the crests (maximal values of pressure) correspond to the center of the compressions whereas the troughs (minimum values of pressure) correspond to the middle of the rarefactions (Figure 11.3).

Like transverse waves, longitudinal waves are

2 Obviously, the air molecules do not stay in fixed positions, but move randomly, so we can consider that they do not have a preferential direction of movement before the generation of the acoustic signal.

characterized by wavelength, frequency, period, amplitude and speed. The distance between two successive crests (or troughs) is the **wavelength**, *λ*, of the acoustic wave and the number of crests (or complete cycles) that pass a given point per second is the **frequency** (**ν**). The **period** (**T**), is the time elapsed between two successive crests passing the same point in space and it equals $1/\nu$. The **amplitude** of the acoustic wave is often defined as peak-to-peak pressure (the difference between the maximum and the minimum pressure) or peak pressure (the difference between the maximum pressure and the baseline pressure). Unless stated otherwise, in this chapter we will consider the *peak pressure* (p_{peak}) as the amplitude of the acoustic wave, which is the amplitude as depicted in Figure 11.3.

If we consider the longitudinal wave in terms of displacement, the periodical oscillation in space and time of the particles of the medium in which the acoustic wave travels is mathematically described by the following equation:

$$y(x,t) = y_{max}\, sin\left[2\pi\left(\nu t - \frac{x}{\lambda}\right)\right] \qquad (11.1)$$

where y is the elongation = the displacement of particles (m); y_{max} is the amplitude (m) = the maximal displacement of a particle from the equilibrium position; x is the spatial coordinate (m); t is the time (s); ν is the frequency (Hz); λ is the wavelength (m).

The wave speed, c, is the speed at which wave crests move forward. A wave crest travels a distance of one wavelength during one period, thus the wave speed is $c = \lambda/T = \lambda\nu$.

1.2. The acoustic pressure

The acoustic pressure, p_a, is the local deviation from the average atmospheric pressure (Figure 11.3B). Its value measured at a certain moment in time is called the *instantaneous pressure*. Compared to the value of the atmospheric pressure ($p_{atm} \approx 10^5$ Pa), the acoustic pressure variations are very small and, in order for the auditory system to detect them, the (baseline) air pressure on the two sides of the tympanic membrane (eardrum) must be equal. In this way, the vibration of the eardrum is related only to the acoustic pressure. The effective pressure[3] exerted onto the eardrum is:

$$p_{eff} = \frac{p_{peak}}{\sqrt{2}} = 0.709\, p_{peak} \qquad (11.2)$$

The human ear can detect acoustic pressures from a minimum p_a of **20 μPa (the audibility limit or hearing threshold)** to a maximum p_a of **20 Pa (the pain threshold).**

1.3. Frequency of the acoustic wave

Acoustic waves have a broad range of frequencies. Depending on the frequency, the acoustic spectrum is divided into several domains:

- *Infrasounds* (frequency < 20 Hz), produced by natural phenomena like earthquakes, thunder, volcanoes or by vibrating heavy machinery;
- *Sounds* (frequency: 20 – 20000 Hz);
- *Ultrasounds* (frequency: 20000 – 10^9 Hz); ultrasonic waves have important medical applications (e.g. ultrasonography, lithotripsy, etc.) that are described in other chapters of this book.

The **human audible range** is from about **20 Hz to 20000 Hz.**[4] The upper end of this range typically decreases with age, and may be as high as 40 kHz in children. Our auditory system translates the information related to frequency of acoustic signal into the sensation of **pitch**. Thus, an acoustic signal having a low frequency evokes a sound having a low pitch (e.g. the sound of a bass drum) and a high frequency induces a high pitch (e.g. the sound of a violin).

Depending on their composition in terms of frequencies, acoustic signals are classified as follows:

- *simple periodic sounds* (pure sounds). These are used clinically for the determination of hearing thresholds;
- *complex periodic sounds* (e.g. speech, music, etc.);
- *random sounds* (e.g. noises).

In order for an acoustic wave to be generated, something must vibrate strongly enough to create an air wave of a certain intensity. For example, if your doctor strikes a tuning fork, you will hear a *pure sound* which is a sinusoidal acoustic wave characterized by a single frequency (Figure 11.4, top panel).

On the other hand, if you pinch a string of a guitar or play a wind instrument, the acoustic signal will be composed of several frequencies, related to each other. Depending on its structure and on its dimensions, each musical instrument has a natural fundamental mode of vibration to which the lowest frequency corresponds. This

3 The *effective pressure* is also known as *root mean square pressure* (or rms pressure). It is calculated by squaring the peak amplitude at each instant, getting an average of the squared values and finally taking the square root of this average. The relationship between p_{eff} and peak pressure presented in this chapter is valid for a sine wave.

4 We can actually hear frequencies a bit lower than 20 Hz, but 20 – 20000 is an easily remembered approximation, and most textbooks give the limits of human hearing as such.

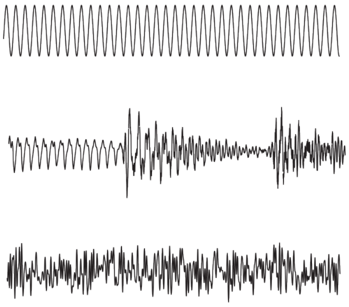


Figure 11.4. Waveforms of different types of acoustic signals. Description is provided in the text.

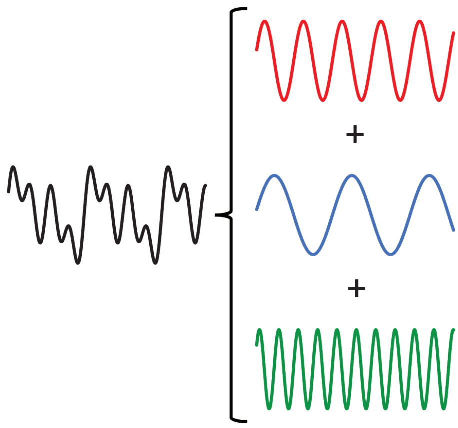


Figure 11.5. An example of a complex periodic sound (left) that is the sum of three harmonic pure sounds (right).

frequency is called the *fundamental frequency* (or *first harmonic*). The other possibilities are integer multiples of the fundamental frequency, called *harmonics* (second, third, etc.). The resulting wave is a combination of the fundamental frequency and various amounts of the harmonics. Although not a pure sinusoid, the wave has a complex repeating pattern and it's classified as a *complex periodic sound* (Figure 11.4, middle panel). The information contained within the combination of frequencies is translated by the auditory system as **timbre** (or quality) of the sound.

Now, suppose you drop a rock on concrete. You will hear a sharp *noise*; an acoustic signal made up of many different frequencies at once, with no relation to each other (Figure 11.4, bottom panel). In this particular case, the noise is also short, because the vibrations of the rock and of the concrete dampen rapidly.

The sounds of the human voice are made up of sequences of periodic sounds (e.g. vowels) and noises sometimes combined with periodic sounds (consonants).

When detecting the complex acoustic waves around us, our ears process the signal in a manner similar to **Fourier analysis**. This is a mathematical technique that allows one to calculate a frequency dependent function from a time-dependent function (like pressure as function of time for the acoustic wave). In other words, using Fourier analysis, we can decompose a complex sound into pure sounds of given frequencies (Figure 11.5). The mechanism will be described later in the chapter as well as in the chapter on Psychophysics.

1.4. The speed of the acoustic wave

The speed of an acoustic wave, ***c***,[5] is not constant and, in general, it depends on the elastic properties of the medium through which it is travelling (described by the bulk modulus, *B*) and the density of the medium (*ρ*).

$$c_{acoustic\ wave} = \sqrt{\frac{B}{\rho}} \tag{11.3}$$

The value of B reflects the stiffness of the medium. In conclusion, the speed increases with increasing stiffness (or decrease in compressibility) and decreases with increasing density of the medium. Thus:

$$c_{in\ solids} > c_{in\ liquids} > c_{in\ gases} \tag{11.4}$$

For gases, the speed of an acoustic wave is given by the Newton-Laplace equation:

$$c_{acoustic\ wave} = \sqrt{\frac{\gamma P}{\rho}} = \sqrt{\frac{\gamma RT}{M}} \tag{11.5}$$

where γ is the adiabatic coefficient, P is the pressure, ρ is the density, R is the ideal gas constant; T is the absolute temperature (K) and M is the molecular weight (kg/kmol).

Equation (11.5) shows that the speed of the acoustic waves in air is higher at warmer temperatures and it also enables us to calculate the speed in certain conditions; for example, the speed of the acoustic signal in air at 300 K (27 °C), considering that M = 28.8 kg/kmol, γ = 1.4, and R = 8310 J/(kmol·K), is:

$$c = \sqrt{\frac{1.4 \cdot 8310 \cdot 300}{28.8}} = 348\frac{m}{s} \tag{11.6}$$

1.5. The acoustic impedance

When an acoustic wave travels, the medium opposes a certain resistance to the propagation of the wave, called *acoustic impedance* (*Z*), which depends on the density of the medium (*ρ*) and the speed of the wave (*c*).

5 For acoustic waves, c rather than v is used as a symbol for the wave speed, in order to distinguish it from the speed of moving objects which are emitting/receiving the acoustic signals. Do not confuse it with the speed of light in vacuum, which has a constant value, even though they are noted with the same abbreviation!

Table 11.1. Speed and acoustic impedance values for medically relevant media.[6]

Medium	c (m/s)	Z_c (10^6 Rayl)
Air	346	0.0004
Fat	1450	1.38
Water (25 °C)	1493	1.48
Soft tissue	1540	1.63
Blood (37 °C)	1570	1.67
Bone	4000	3.8 – 7.4

$$Z = \rho c \tag{11.7}$$

The unit for Z is Rayl (Rayleigh),[7] also called acoustic ohm. Under well-defined pressure and temperature conditions, the impedance of a medium of density ρ_0 is called the characteristic impedance and is denoted by Z_c:

$$Z_c = \rho_0 c \tag{11.8}$$

At a temperature of 20 °C and an atmospheric pressure of 1 bar ($1.0131 \cdot 10^5$ Pa), the density of air is $\rho_0 = 1.206$ kg/m^3 and the speed of acoustic waves in air is 343 m/s, so the characteristic impedance of air is $Z_c = 413$ Rayl (in clinical audiology it is approximated to 400 Rayl).

Acoustic impedance values in different media generally fall within the following range of values (in Rayls):

- $10^2 - 10^3$ for gases;
- 10^6 for liquids;
- 10^7 for solids.

Therefore:

$$Z_{gas} < Z_{liquid} < Z_{solid} \tag{11.9}$$

Table 11.1 shows some acoustic impedance values of interest in medicine.

Acoustic impedance is a crucial parameter in determining how waves will reflect and transmit at boundaries between media. If we considered the incident wave propagating in medium 1, then the fraction of the incident signal that is transmitted in medium 2 or reflected back to medium 1 depends on the impedances of the two media (Z_1 and Z_2, respectively). If the impedance is matched (the impedances are equal), the acoustic signal crosses the boundary without reflection. If they are different, some of the incident signal will be reflected (echo). The strength of the echo produced as acoustic waves is expressed by the reflection coefficient (R), which, for an incident wave perpendicular to the boundary between two media is:

$$R = \frac{(Z_1 - Z_2)^2}{(Z_1 + Z_2)^2} \tag{11.10}$$

In conclusion, **the greater the difference in acoustic impedance, the stronger the reflection and the smaller the transmission.** Keep this in mind for later, when we will discuss about the role of the middle ear. This concept is also important for understanding the principle of *ultrasonography* (see the chapter on Medical imaging).

Let us consider an acoustic wave propagating through air and reaching the surface of water. Knowing that $Z_{air} = 413$ Rayl and $Z_{water} = 1.44 \cdot 10^6$ Rayl, we can calculate the proportion of the incident wave that would be reflected using equation (11.10) and we obtain R = 0.9988. The proportion of the incident signal that is transmitted in these conditions would be $(1 - R) \approx 0.001$.

Thus, in case you were wondering why sounds seem to fade away when you submerge your head in water (for example when diving), the answer is because only about one thousandth of the sound wave in the air is transmitted into the water.

1.6. Intensity of the acoustic wave

Waves transport energy from one place to another. For a sinusoidal wave, it has been calculated that **the energy transmitted by the wave is proportional to the square of the amplitude of the wave**. There are two parameters that quantify the energy transport:

- **power** is a measure of how much energy is transmitted per time and is measured in W (watt);
- **intensity**, I, is a measure of how much power is transported across unit area perpendicular to the direction of energy flow. It is measured in W/m^2, as:

$$I = \frac{energy/time}{area} = \frac{power}{area} \tag{11.11}$$

The intensity of the acoustic wave can be also expressed in terms of acoustic pressure and impedance:

$$I = \frac{{p_a}^2}{Z} \tag{11.12}$$

Considering the minimum audible p_a of 20 µPa and the air acoustic impedance of 400 Rayl, we can calculate the minimum intensity of the acoustic wave that the human ear can detect:

$$I = \frac{(20 \cdot 10^{-6})^2}{4 \cdot 10^2} = 10^{-12}\ W/m^2 \tag{11.13}$$

Thus, the human acoustic sensitivity threshold is 1 pW/m^2 and, if we calculate the intensity for a

6 Data according to Suetens, P. (2009). *Fundamentals of Medical Imaging*: Cambridge University Press.

7 1 Rayl = 1 kg · s^{-1} · m^{-2}.

maximum p_a, the maximum intensity of 1 W/m^2 will be obtained (pain threshold).

1.7. Sound intensity level. The decibel scale

The intensity of the acoustic wave induces the sensation of **loudness** of a sound. Remember that the intensity of the acoustic wave is directly proportional to the square of the acoustic pressure, as described by equation (11.12). Thus, the property of the acoustic wave directly related to the sound loudness is the amplitude of the pressure variation. For more details regarding the relation between the physical properties of the acoustic wave and the subjective properties of sound (pitch, timbre, loudness), see the Psychophysics chapter.

We saw in the previous section that an average human ear can detect acoustic signals with an intensity range spanning a factor of 10^{12} from lowest to highest. However, what we perceive as loudness is not directly proportional the intensity. Experiments have showed that, in order to produce a sound A that is perceived twice as loud as the sound B, the intensity of the acoustic wave A should be 10 times higher than the intensity of the wave B. For example, for average persons, an acoustic signal of intensity 10^{-2} W/m^2 sounds twice as loud as one with intensity 10^{-3} W/m^2 and four times as loud as 10^{-4} W/m^2. This is valid for frequencies near the middle range of the audible domain.

This relationship between the intensity (a physically measurable quantity) and loudness (a subjective sensation) can be represented using a logarithmic scale. The unit on this scale is **bel (B)**, but **decibel (dB)** is much more commonly used; 1 dB = 0.1 bel. In general, a value expressed in decibels specifies a ratio, in this case, between the intensity of a certain acoustic signal, I, and a reference intensity, I_0, ($I_0 = 10^{-12}$ W/m^2). Therefore, the **sound intensity level,** L_I, is:

$$L_I = 10log_{10}\frac{I}{I_0} \qquad (11.14)$$

In a similar manner, a **sound pressure level**, L_p, can be defined, also in dB, by taking the minimum detectable acoustic pressure, p_0, as a reference (p_0 = 20 µPa = 2 x 10^{-5} Pa):

$$L_p = 20log_{10}\frac{p}{p_0} \qquad (11.15)$$

The values of I_0 and p_0 correspond to the normal acoustic sensitivity threshold at the frequency of 1000 Hz.

Considering the expression of intensity given by equation (11.12), we can rewrite equation (11.14) as follows:

$$L_I = 10log_{10}\frac{I}{I_0} = 10log_{10}\left(\frac{p}{p_0}\right)^2 = 20log_{10}\frac{p}{p_0} = L_p \qquad (11.16)$$

In conclusion, **the sound intensity level and the sound pressure level have the same value for the same acoustic signal.**

We know from experience that the loudness of a sound decreases as we move further away from the source. This is because the intensity decreases with the square of the distance. This effect is easily observed if a source is in the open and the acoustic waves can freely radiate in all directions. Indoor, because of reflections from the walls, this might not be true anymore.

The intensity decreases due to dissipation of energy as heat when particles vibrate. The loss of intensity happens faster for higher frequencies (energy dissipates faster) than for lower ones. This is the reason why, at an outdoor concert, when you are far from the stage, you hear manly the low frequencies emitted by the boom of the drums.

2. BIOPHYSICS OF SOUND RECEPTION

2.1. Auditory system. Introductory notions

The sensation of sound is the result of a complex chain of events that involve:

- detection, amplifying and processing of the acoustic signal;
- conversion of the acoustic signal into changes of membrane potential of the sensory cells;
- generation of action potentials in neurons from the auditory branch of the vestibulocochlear nerve (cranial nerve VIII);
- transmission of these action potentials towards the auditory cortex and finally, their integration in the temporal lobe.

All these tasks are performed by structures of the auditory system which is formally divided into:

- the peripheral auditory system;
- the central auditory system.

The peripheral auditory system consists of the outer ear, middle ear and inner ear (Figure 11.6) and it performs three main functions:

- **transmission of the acoustic signal** is carried out in two ways: *the air way*, performed by the outer ear and the middle ear and *the bone way*, which has a secondary role in natural hearing, but which is of great interest in clinical and audiological practice;
- **analysis and amplification of the acoustic signal** is provided by the different compartments of the inner ear and involves complex passive and active mechanisms, starting with a Fourier analysis at the level of the basilar membrane and

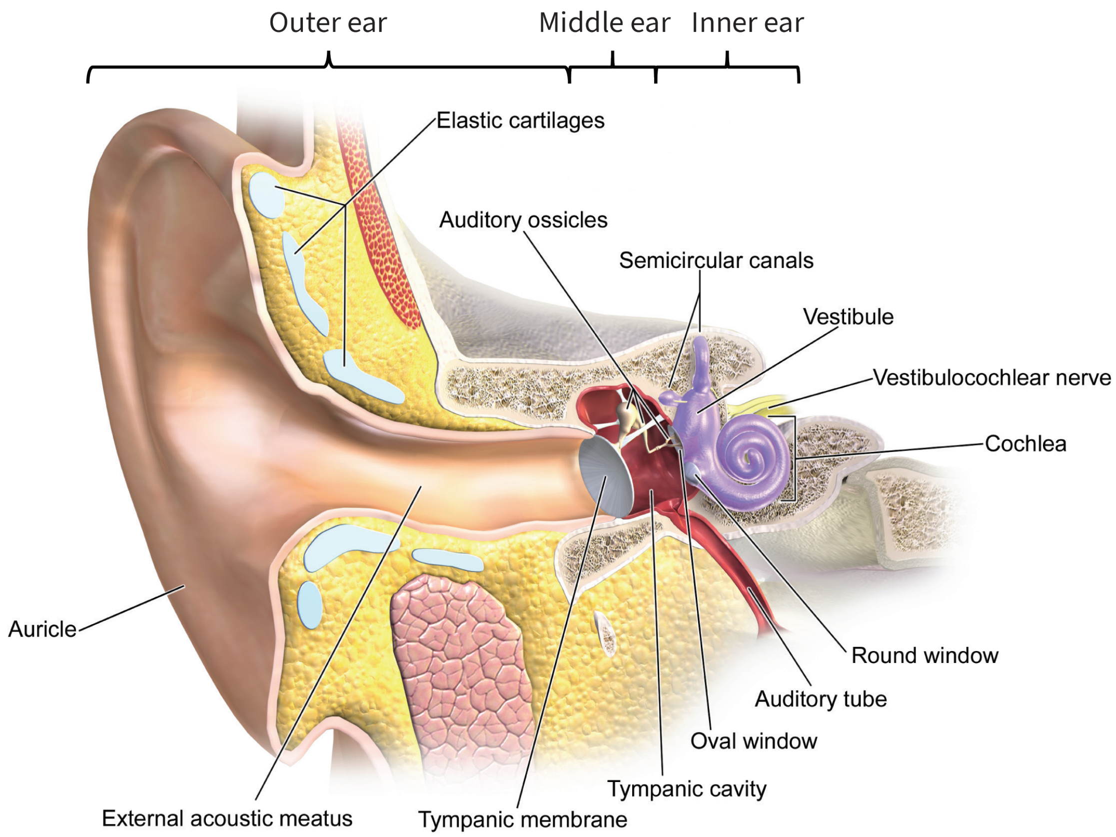


Figure 11.6. Anatomy of the peripheral auditory system.[8]

ending with the active amplification of the signal at the level of the outer hair cells of the organ of Corti;

▶ **translation of the acoustic signal into electric signals (mechanotransduction)** is performed by the hair cells. The electric signals are then relayed via the auditory branch of the vestibulocochlear nerve to the central auditory system.

The *central auditory system* begins as the cochlear (auditory) branch of the vestibulocochlear nerve, ascends from the cochlea to higher order auditory structures and functions to process the output from the peripheral auditory system. Due to the activity of central auditory system, we get the sensation of sound, we can discriminate among sounds and localize their sources.

In the rest of this chapter, we will focus on the peripheral structures and the biophysical processes that take place at this level.

2.2. The outer ear

As Figure 11.6 shows, proceeding form the outside to inside, the most visible part of the outer ear is the *auricle* (or *pinna*), a skin covered cartilage structure which funnels the acoustic waves into the *external auditory canal* (also called *external acoustic meatus*) in order to focus them on the *tympanic membrane* (or the *eardrum*).

The **auricles** are very important for the localization of the sound sources in the vertical plane. Acoustic signals can enter the auditory external canal both directly and after being reflected by the auricle and our sensation of sound is triggered by the combination of the two. Depending on their angles of incidence, the acoustic waves are reflected differently off the auricle and its small extension, the *tragus*, therefore, a signal coming from above our head will sound differently than a signal coming from straight in front of us. In addition to that, due to the shape of the auricle, and especially the *concha* (a hollow depression next to the external auditory canal), certain acoustic frequencies are emphasized over others.

The **external auditory canal** penetrates about 2.5 cm into the temporal bone, it has an average diameter of about 5 to 10 mm and it ends blindly at the **tympanic membrane** which vibrates, much like the head of a drum, in response to the acoustic waves reaching it. Acoustically, the tympanic membrane is the final part of the outer ear, which thus functions as a tube filled with air open only at one end, amplifying the waves whose frequency

8 Modified from a copyright-free image provided by Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". WikiJournal of Medicine 1 (2). DOI:*10.15347/wjm/2014.010*. ISSN 2002-4436.

matches one of its own natural frequencies of vibration, a phenomenon called *acoustic resonance*.[9] In general, the resonant frequency of vibration, v, of a musical instrument, the length of the vibrating string or pipe, L, and the speed of the acoustic wave, c, are mathematically corelated. For a tube open only at one end, the corelation is:

$$v_n = n\frac{c}{4L} \qquad (11.17)$$

where n = 1, 3, 5, etc.

Considering the length of the external auditory ear canal of 2.5 cm and the speed of the acoustic waves in air of 343 m/s, the resonant frequencies v_n, can be is easily calculated, for example:

$$v_1 = \frac{343\,\frac{m}{s}}{4 \cdot 0.025\,m} = 34\,300\,Hz = 3.43\,kHz \qquad (11.18)$$

In conclusion, the functions of the outer ear are:

- **directing and focusing the acoustic signals onto the tympanic membrane**: the auricle behaves like an acoustic antenna;
- **emphasizing certain frequencies over others**: the concha and the external auditory canal act as a resonator;
- **helping in localization of the source of the acoustic signal**.

2.3. The middle ear

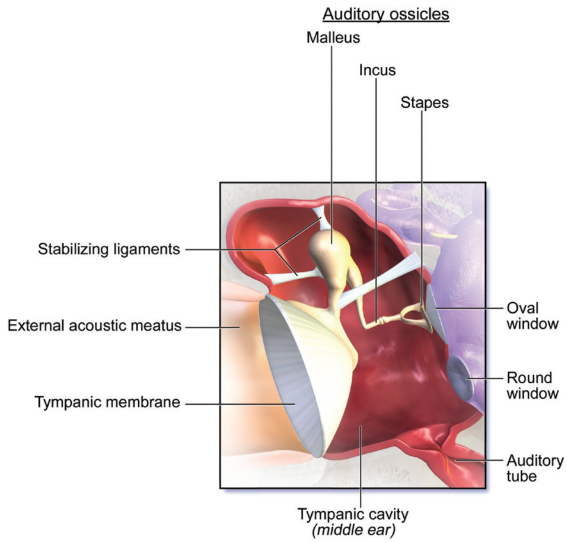


Figure 11.7. Anatomy of the middle ear.[10]

The middle ear (Figure 11.7) is an air-filled cavity in the temporal bone communicating with the outer ear via the tympanic membrane and with the inner ear via two other membranes covering two orifices called the **oval window** and the **round window**. Inside the middle ear there are three small bones (known collectively as ossicles) suspended by ligaments. One of the ossicles, the **malleus** (commonly called the *hammer*) is attached to the tympanic membrane. The malleus is connected to the **incus** (or *anvil*), which in turn connects to the **stapes** (often called the *stirrup*), which is attached to the oval window.

The middle ear is connected to the oral cavity through the *auditory tube* (also called the *Eustachian tube*), an important structure for equalizing the air pressure on opposite sides of the tympanic membrane. Remember that the acoustic pressure values are quite small relative to atmospheric pressure, so, in order for us to be able to detect them, the air pressure on both sides of the tympanic membrane must be the same.

When you experience sudden changes in altitude (e.g. in an ascending or descending elevator or airplane), the air pressure outside and in your outer ear changes relative to the middle ear. Because of that, the tympanic membrane is stretched and you may experience an unpleasant sensation of clogged ears. The pressure in the middle ear stays initially the same because the auditory tube is closed. However, voluntarily swallowing or yawning open the auditory tube and allow air to flow between the middle ear and the oral cavity, and therefore to equalize the pressure on the two sides of the eardrum.

The primary function of the middle ear is to transfer vibrations of the tympanic membrane to the oval window in order to enter the inner ear, which, as we will see later, is filled with liquid. Thus, the acoustic signals travel through air, enter the outer ear, reach the tympanic membrane and cause its vibration and, from there, they must be transmitted to the liquid in the inner ear.

If you remember our discussion about the acoustic impedance in the beginning of this chapter, you will probably ask how is it possible for the acoustic signal to enter the fluid in the inner ear, knowing that the acoustic impedance of a liquid (e.g. water) is much greater than the acoustic impedance of air? Wouldn't most of the acoustic waves be reflected at the oval window as equation (11.10) predicts? The answer is no, because there are some features of the ear's anatomy that help.

First, the ossicles act like a system of levers and, through their movement, they slightly amplify

9 To be more specific, the frequency that is amplified by acoustic resonance is that of a *standing wave*, i.e. the result of interference of the incoming and reflected acoustic waves (e.g. by the tympanic membrane) having the same frequency. The term "standing" is due to the fact that it oscillates in time, but its wavelength does not change, and neither does the location of the peaks and throughs.

10 Modified from a copyright-free image provided by Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". WikiJournal of Medicine 1 (2). DOI:*10.15347/wjm/2014.010*. ISSN 2002-4436.

the vibrations by a factor of around 1.3. Second, and most important, there is a difference in the surface area between the tympanic membrane (effective area ~55 mm^2) and the oval window (~3.2 mm^2) and this has a more dramatic effect. The force applied on the ossicles is the product of the area of the tympanic membrane and the pressure exerted on it.[11] This force is amplified by the movement of the ossicles and then applied to a much smaller area, the oval window. If we calculate the ratio between the two areas given above, we will conclude that the pressure acting on the much smaller surface of the oval window is about 17.2 times higher. In other words, due to different areas of the two membranes, the pressure is amplified by a factor of 17.2. If we also take into account the amplification factor of the ossicles, we will obtain an overall amplification factor of about 22 (1.3 · 17.2 = 22.4).

Due to its ability to amplify the pressure as it is transmitted from the outer ear to the inner ear in order to overcome the difference in impedance, the middle ear functions as an *impedance matching device*. **In this way, rather than being reflected, most of the energy of the acoustic signal is transferred to the liquids of the inner ear**.

Damage to the outer or middle ear lowers the efficiency of the acoustic signal transmission to the inner ear. This medical condition, called *conductive hearing loss*, can be partially overcome by artificially boosting sound pressure levels with an external hearing aid.

The middle ear is also able to reduce the transmission of the acoustic signal into the inner ear, in order to lower the risk of damage of the delicate hair cells from the inner ear. The ossicles are suspended by ligaments connected to muscles that can reduce their movement. The tensor tympani and the stapedius muscles are the smallest muscles in the human body and, through their synchronized contractions, they can limit the movement of the malleus and stapes, respectively. The two muscles are part of the *acoustic reflex* (also called *stapedius reflex)*. However, this protection mechanism is not very fast, there is a time lag of a few milliseconds and the stapedius muscle cannot withstand a prolonged contraction. Therefore, if you are exposed to loud, abrupt noise, the acoustic reflex will offer poor protection and your experience is more likely to be painful and even damaging. On the other hand, there are conditions (such as Bell's palsy) associated with a reduction in

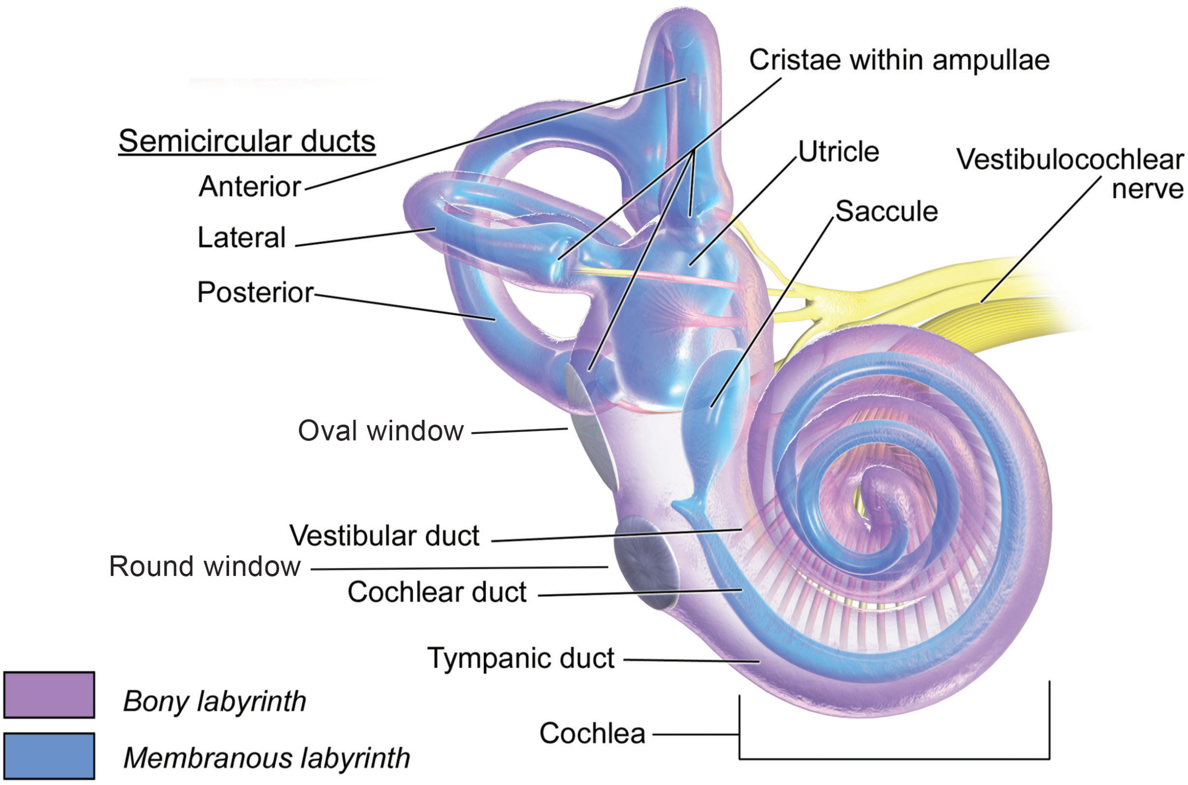


Figure 11.8. Anatomy of the inner ear.[12]

11 Pressure = force/area, therefore force = pressure · area.

12 Modified from a copyright-free image provided by Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". WikiJournal of Medicine 1 (2). DOI:*10.15347/wjm/2014.010*. ISSN 2002-4436.

the muscle tone of either of these muscles which leads to *hyperacusis*, a painful sensitivity to mild or even low intensity acoustic signals.

2.4. The inner ear

The components of the inner ear (Figure 11.8) are placed in a complex system of cavities, within the temporal bone, collectively called the *bony labyrinth*, which has three parts:

- *three semi-circular ducts* (important for the sense of equilibrium);
- the *cochlea*[13] (important for the sense of hearing);
- the *vestibule* which connects the former two parts.

The bony labyrinth contains a seawater-like fluid called *perilymph.* Inside the bony labyrinth there is the *membranous labyrinth*, which consists of the *semi-circular ducts*, *utricle* and *saccule* (inside the vestibule) and the *cochlear duct*, and it is filled with another type of fluid called *endolymph*.

The auditory portion of the inner ear is represented by the **cochlea**, a tubular structure of about 3.5 cm long and coiled 2.5 times around a central bony column called the columella. The cochlea is almost entirely divided lengthwise from its base by the **cochlear duct** (or *scala media*), an endolymph-filled membranous tube, which follows the cochlear spiral and contains the sensory cells of the auditory system. As we will see later, these receptor cells send messages to the neurons composing the cochlear branch of the vestibulocochlear nerve. On either side of the cochlear duct there are two perilymph-filled compartments: the **vestibular duct** (or *scala vestibuli*), which is on the side of the cochlear duct that begins at the oval window, and the **tympanic duct** (or *scala tympani*), which is below the cochlear duct and ends at the round window. The vestibular and tympanic ducts meet at the end of the cochlear duct at the **helicotrema.** The oval and the round windows are at the basis of the cochlea, while the helicotrema is at its apex (Figure 11.9).

How are these structures involved in hearing? Let's summarize what we know so far. The acoustic signal enters the external auditory canal and triggers the vibration in and out of the tympanic membrane which moves the middle ear ossicle chain against the oval window amplifying the acoustic pressure. Therefore, the oval window is the entrance of the acoustic signal into the cochlea. The movement of the stapes creates pressure waves into the vestibular duct, having the same frequency as the frequency of vibration of the oval window, which spread out very fast. A great part of the vestibular duct wall is bone, and there are only two ways by which the pressure wave can dissipate. One way is to travel along the whole vestibular duct and pass into the tympanic duct through the helicotrema. However, most of the pressure is transmitted from the vestibular duct across the cochlear duct. Because the fluids in the cochlea are incompressible, the pressure waves created in the tympanic duct are relieved by the vibration of the round window at the end of the tympanic duct.

As illustrated in Figure 11.10, a cross-section through the uncoiled cochlea reveals three sections: the vestibular and the tympanic ducts filled with perilymph and the cochlear duct with endolymph. The perilymph has an ionic composition similar to the cerebrospinal fluid (low in K^+ and high in Na^+). The perilymph from the

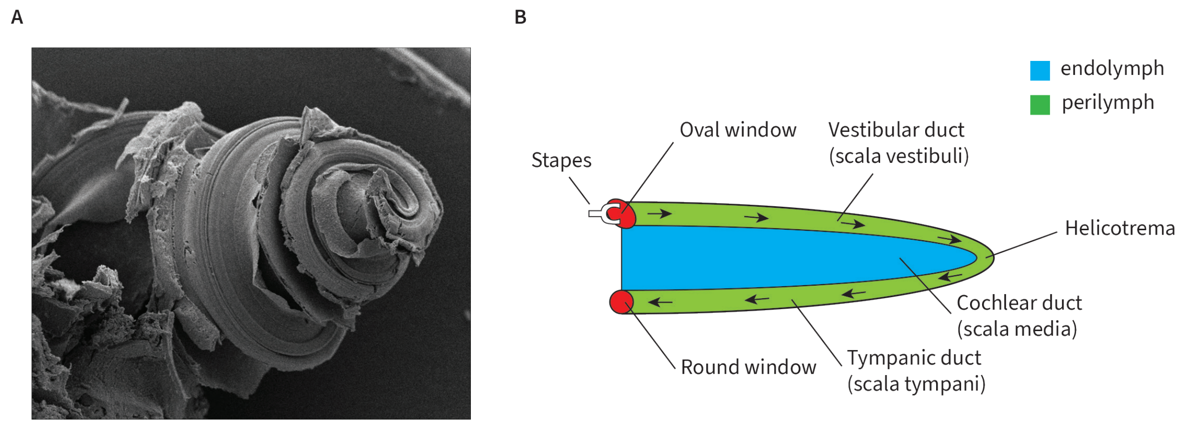


Figure 11.9. The cochlea. A, scanning electron micrograph[14] of a cochlea removed from the bony labyrinth. B, schematic representation of an uncoiled cochlea. The vestibular and tympanic ducts are filled with perilymph, and the cochlear duct is filled with endolymph. The vestibular duct starts at the oval window and meets the tympanic duct at the helicotrema. The tympanic duct ends at the round window. Arrows show the direction of travel of the acoustic wave.

13 *Cochlea* is the Latin word for snail.

14 Public domain image from Steph Hares, University of Bristol. CC0 1.0 Universal. Source: Wellcome Collection. *https://wellcomecollection.org/works/wyz53ctv*

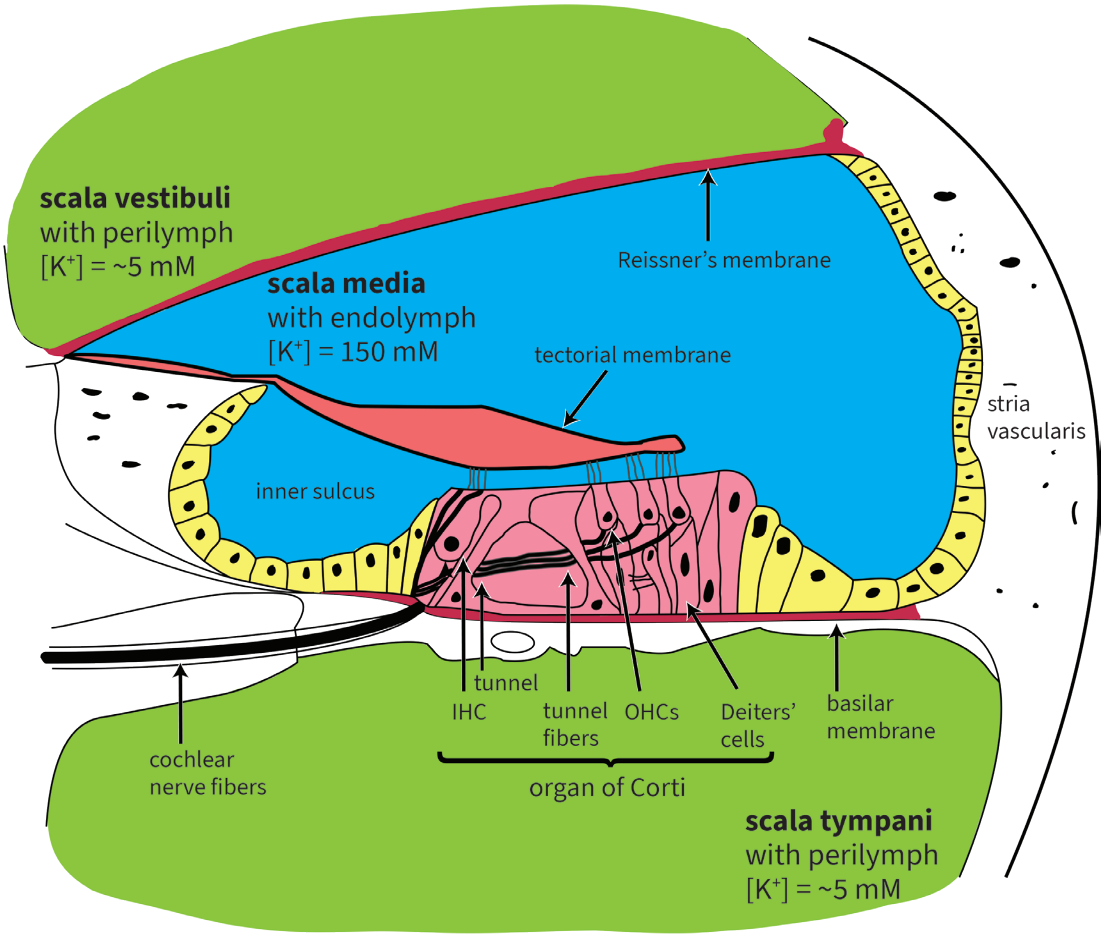


Figure 11.10. Cross-section through the cochlea.[15] Description is given in the text. The K^+ concentrations of endolymph and perilymph are also given. IHC = inner hair cell, OHCs = outer hair cells.

vestibular and tympanic ducts communicates at the helicotrema. By contrast, the endolymph is extremely rich in K^+ and low in Na^+, thus more like the cytoplasm. Due to the K^+ concentration gradient, there is a difference in electric potential between the endolymph and perilymph of about +80 mV, called the **endocochlear potential**. A frequent cause for hearing loss is the reduction in endocochlear potential, proving its critical role in sensory transduction, as we will discuss below.

Upon the basilar membrane sits the **organ of Corti**, containing the sound sensory cells, called **hair cells**. The organ of Corti stretches the length of the basilar membrane and consists of (Figure 11.11) one row of **inner hair cells** (~3500) and three rows of **outer hair cells** (~16000). These are epithelial cells named after the bundle of hairlike structures called stereocilia[16] on their apical end protruding into the cochlear duct. The hair cells lie in a matrix of supporting cells and their stereocilia are in contact with a gel-like structure, called the **tectorial membrane**.[17] The tectorial membrane is attached only along one edge, with a kind of hinge, the other end being free into the endolymph. Note that mature hair cells do not divide, thus new cells cannot be formed to replace dead cells, making hearing loss irreversible.

2.5. The cochlear hair cells as sensory cells

At this point we know where the sensory cells are, but how do they detect the acoustic signal? The stereocilia of the inner hair cells have a dense oblique cone-like arrangement, while those of

15 Modified from an image by Oarih (*https://commons.wikimedia.org/wiki/File:Cochlea-crosssection.svg*) available under a Creative Commons license (*https://creativecommons.org/licenses/by-sa/3.0/deed.en*).

16 The cilia present on the apical surfaces of many types of epithelial cells contain a specific arrangement of microtubules. Those present on the auditory hair cells have a different structure, and for this reason they are called *stereocilia*, but the term *stereovilli* is also used.

17 Until recently, only the outer hair cells were believed to be in direct contact with the tectorial membrane at rest, but a very recent study (Hakizimana, P., & Fridberger, A. (2021). Inner hair cell stereocilia are embedded in the tectorial membrane. *Nature Communications*, 12(1), 2604. doi:10.1038/s41467-021-22870-1) suggests that this is valid also for the stereocilia of the inner hair cells.

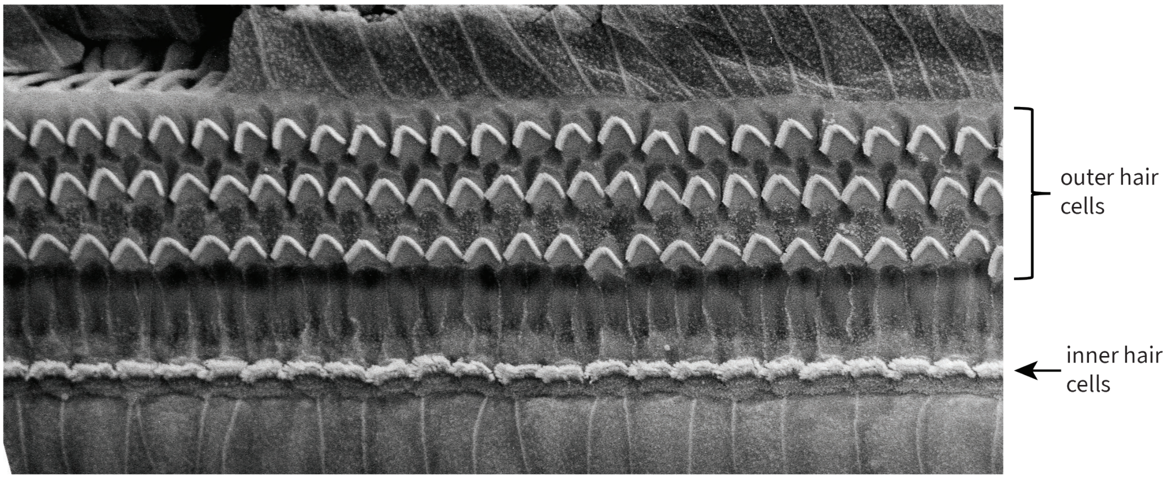


Figure 11.11. Scanning electron micrograph of the guinea pig organ of Corti.[18] The organ of Corti contains 3 rows of outer hair cells and 1 row of inner hair cells.

the outer hair cells are arranged in a "w" shape (Figure 11.11, Figure 11.12).

At the insertion point into the apical membrane, each stereocilium narrows forming a hinge about which it can swing. There are from 30 to a few hundred stereocilia in a hair bundle, they are graded in height and with a bilaterally symmetric arrangement. Vibrations of the basilar membrane cause movement of the hair cells relative to the tectorial membrane and, as a consequence of that, a shear force is generated at their apical end (for example, due to flow of endolymph in or out of the inner sulcus), displacing the stereocilia (Figure 11.13). The appropriate stimulus for activating a hair cell is the movement of its stereocilia, but not just any deflection will do. **Movement of the hair bundle toward the tallest stereocilia depolarizes the hair cell, while displacement toward the shortest hyperpolarizes the cell**. How is that possible?

The epithelium containing the hair cells separates perilymph from endolymph. The basolateral side of the hair cells is bathed in the K^+- poor perilymph, while the K^+- rich endolymph bathes the apical side, including the stereocilia (Figure 11.13). Parallel to the plane of bilateral symmetry of the cochlea, the adjacent stereocilia are connected via filamentous structures known as *tip links* which, when stretched, directly open cation-selective ion channels called mechano-electrical transduction (MET) channels (Figure 11.14). These are permeable to K^+ (but also to Ca^{2+}). Additionally, *lateral links* (*side links*) also exist connecting stereocilia of the same row or of different rows.

At rest, the MET channels are partly open, but when the hair bundle is displaced toward the tallest stereocilium, the channels open completely, and mainly K^+ ions enter the cell. If the hair bundle moves toward the shortest stereocilia, the apical MET channels close.

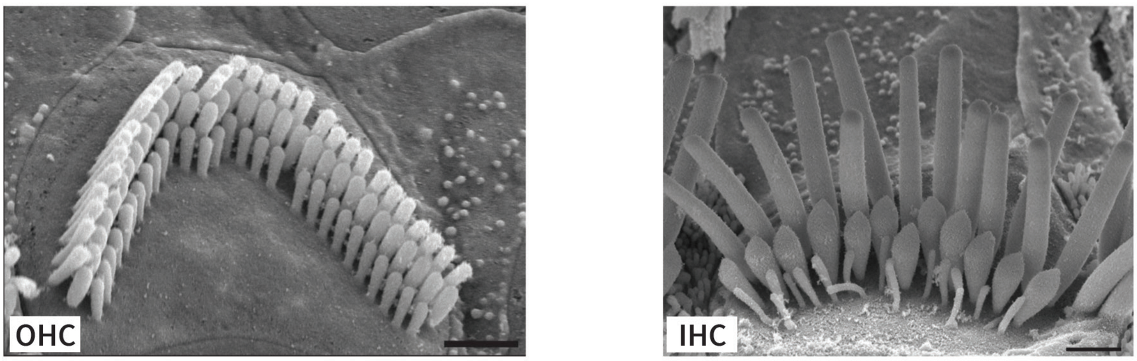


Figure 11.12. Scanning electron micrographs of the stereocilia of hair cells.[19] Left – outer hair cells (OHC). Right – inner hair cells (IHC). Scale bars are 1 µm long in both panels.

18 Image modified from: SEM organ of Corti, rows of hair cells. Prof. Andrew Forge. Attribution 4.0 International (CC BY 4.0). Source: Wellcome Collection. *https://wellcomecollection.org/works/gwqjj9qg*

19 Modified from Ivanchenko, M. V., Indzhykulian, A. A., & Corey, D. P. (2021). Electron Microscopy Techniques for Investigating Structure and Composition of Hair-Cell Stereociliary Bundles. Frontiers in Cell and Developmental Biology, 9. doi:10.3389/fcell.2021.744248, available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*).

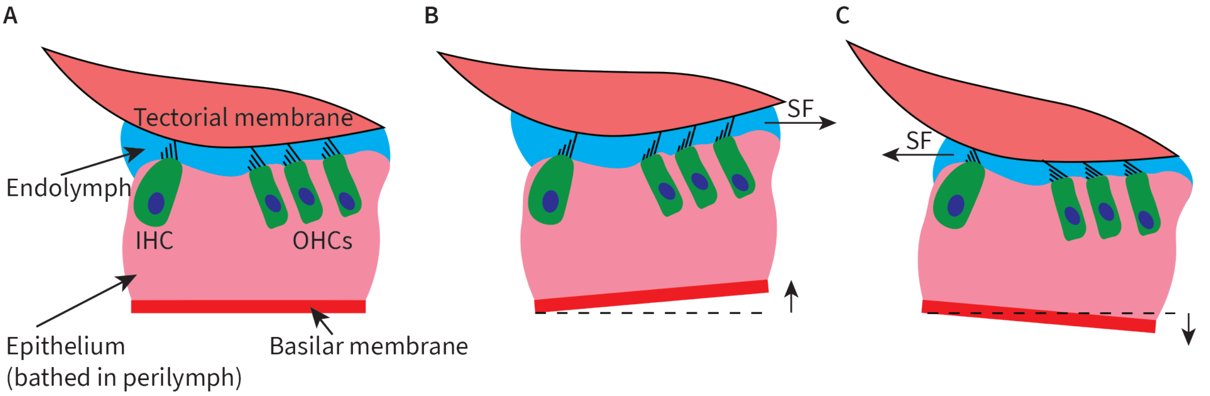


Figure 11.13. Vibration of the basilar membrane causes movement of the hair cells relative to the tectorial membrane and a shear force is generated at their apical end. A, resting position. B, upward movement of the basilar membrane displaces the stereocilia toward the tallest (up). C, downward movement of the basilar membrane displaces the stereocilia toward the shortest (down). IHC = inner hair cell, OHCs = outer hair cells, SF = shear force (see the chapter on Fluid dynamics and hemodynamics).

Remember that between the endolymph and the perilymph there is a difference in electric potential of about +80 mV, therefore the resting membrane potential of the hair cells can be defined in two ways, depending on the reference point: an apical resting potential of about −125 mV relative to endolymph and a basolateral resting potential of about −45 mV relative to perilymph. Because the stereocilia protrude into the endolymph, the electric potential gradient across the membrane of the stereocilia is about 125 mV and this is what drives K^+ entry into the hair cell (even though the cytoplasmic K^+ concentration is already high) and leads to depolarization of the hair cell membrane.

Both types of hair cells depolarize in response to the stereocilia movement in the direction of the tallest, however, the events following depolarization are different between the two, hence the different roles which are attributed to the inner and outer hair cells. The inner hair cells are the sensory cells that convert the acoustic signal into electric signal and about 95% of the fibers of the cochlear branch of cranial nerve VIII receive inputs from these cells. The outer hair cells communicate mostly with efferent neurons (e.g. coming from the superior olivary complex) and the currently available data suggest that they are important for cochlear amplification of the acoustic signal, by modulating the vibration of the basilar membrane (see below).

If the cochlear cells depolarize due to K^+ ions entry through the mechanosensitive channels, how do they repolarize? The repolarization is initiated at the basolateral end, due to K^+ exit through mechano-insensitive K^+ channels down the K^+

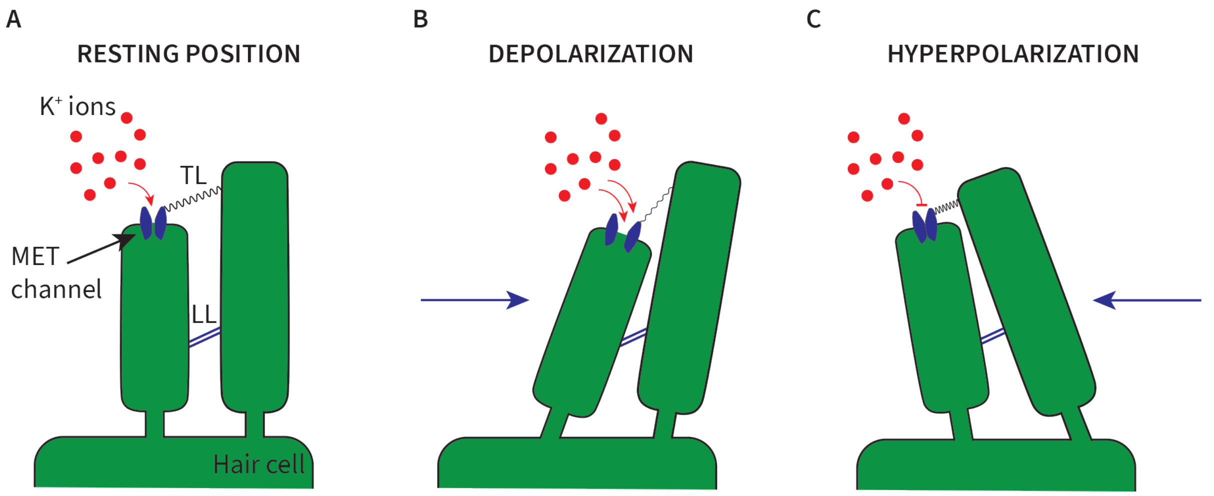


Figure 11.14. Position of stereocilia controls K^+ entry into the hair cell. A, when the hair cell is at rest, the mechanosensitive MET channels at the apical end of the hair cell are partially open, allowing only small amounts of K^+ to enter the cell. B, C, the swing of stereocilia from side to side changes the tension in the tip links modulating the ions flow: movement toward the tallest stereocilium fully opens MET channels, allowing more K^+ to flow inside the cell (B), while movement in the opposite direction closes them (C). As a result, a graded receptor potential is generated correlated to the movement of stereocilia: depolarization (B) or hyperpolarization (C). Horizontal arrows indicate direction of movement of the stereocilia. TL = tip links, LL = lateral links.

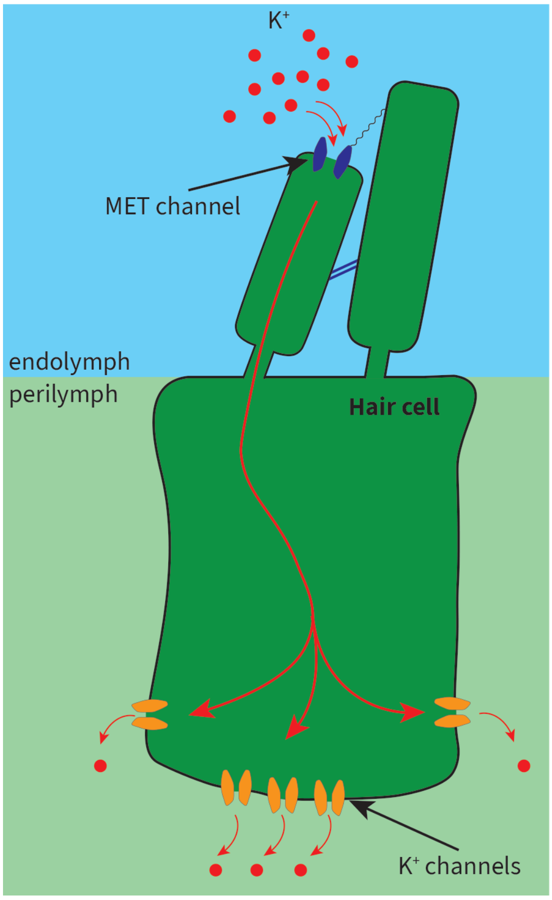


Figure 11.15. Potassium flow through the hair cells. Potassium enters the hair cells at the apical end, through mechano-sensitive MET channels in the stereocilia and exits at the basolateral side of the hair cell, through K^+ channels (that are not mechano-sensitive). The movement direction of K^+ is controlled by the different compositions of the endolymph (K^+-rich) and perilymph (K^+-poor), that bathe the different sides of the hair cell.

electrochemical gradient; this part of the hair cell is bathed in perilymph, which is low in K^+ and the equilibrium potential for K^+ is more negative than the hair cells basolateral resting potential. Thus, K^+ ions enter the hair cell at the apical end and leave it at the basolateral end (Figure 11.15).

When the hair bundle moves toward the shortest stereocilia, the apical K^+ channels close and, because K^+ efflux at the basolateral end continues, the membrane of the hair cell hyperpolarizes (Figure 11.14C).

2.6. The basilar membrane as an acoustic analyzer

In the first part of this chapter, we talked about the properties of the acoustic signals that our ears detect and which, in most of the cases, are a complex combination of frequencies. So, a reasonable question would be – how does the basilar membrane vibrate in response to complex acoustic signals?

For the moment, let's try to understand how the movement of the stapes on the oval window leads to movement of the basilar membrane by analyzing the outcome of the stapes pushing and pulling the oval window. When the stapes pushes the oval window, it causes it to bow toward the vestibular duct. The pressure in the vestibular duct will be higher than the pressure in the tympanic duct and the basilar membrane will bow toward the tympanic duct (downward, as shown in Figure 11.16). When the oval window bows toward the middle ear, the pressure in the vestibular duct falls below the pressure in the tympanic duct and the basilar membrane will bow toward the vestibular duct (Figure 11.16).

Experiments have shown that the vibration of different parts of the basilar membrane and the electrical activity of the cochlear neurons communicating with the hair cells from those parts are tuned, meaning that although the neurons respond to broad range of vibration frequencies, they respond most intensely to a specific frequency of basilar membrane vibration.

Over time, various hypotheses were meant to explain the frequency tuning within the cochlea. According to the currently accepted theory, the

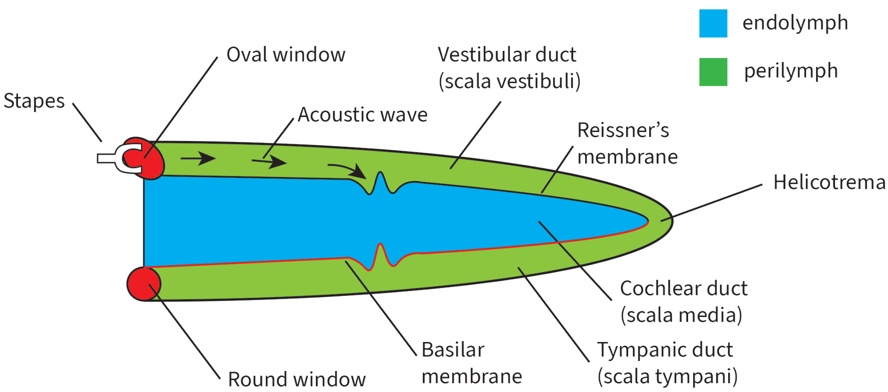


Figure 11.16. Vibration of the basilar membrane in the uncoiled cochlea, based on Figure 11.9B. Displacement of the basilar membrane occurs due to the pressure gradient between the vestibular and the tympanic ducts. Note that Reissner's membrane also bows due to the pressure gradient. However, only the vibration of the basilar membrane activates hair cells.

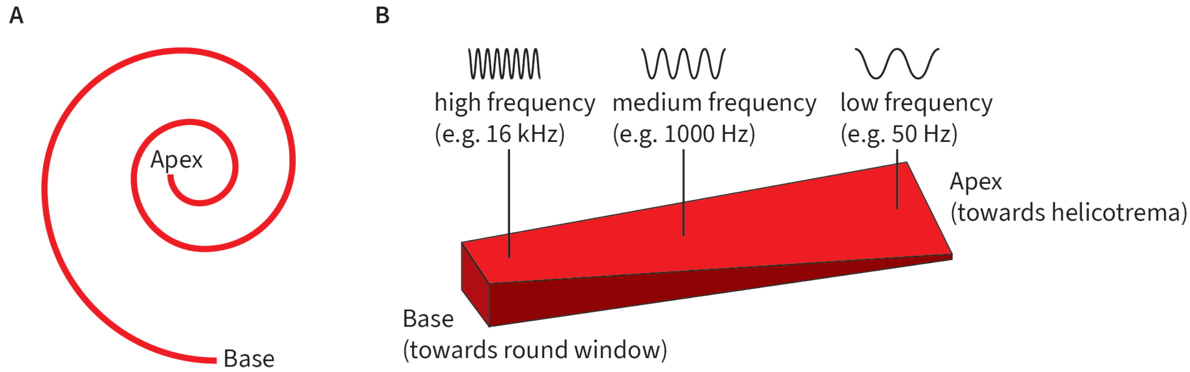


Figure 11.17. Tonotopic map of the basilar membrane. A, two-dimensional view of the basilar membrane coiled in the cochlea. B, three-dimensional view of the uncoiled basilar membrane. Different areas of the basilar membrane vibrate at different frequencies – high frequencies at the base and low frequencies at the apex.

ability of the basilar membrane to vibrate in response to certain frequency of the acoustic wave is partly explained by its geometry and viscoelastic properties: **narrower and stiffer at its basal end and wider and more flexible at its apical end**. This theory is based on the experiments performed by Georg von Békésy[20] while working at the Harvard University on cadaver ears, which have shown that a membrane with variable width and flexibility vibrates maximally at different positions in response to acoustic signals having different frequencies. In addition to that, von Békésy found that a pure acoustic wave triggers a **traveling wave** in the basilar membrane that propagates from its base toward the apex increasing in amplitude and slowing in velocity until a site of maximal displacement is reached. In other words, *the site of maximal vibration depends on the frequency of the acoustic signal*. High frequency acoustic signals induce vibrations at the base of the basilar membrane, whereas lower frequencies cause vibrations toward the apex. In this way, the basilar membrane performs a topographical mapping of frequencies (*tonotopy*), as illustrated in Figure 11.17. A complex acoustic signal induces a pattern of basilar membrane vibrations equivalent to the superposition of the vibrations generated by individual frequencies making up the complex acoustic signal.

Then, to answer the question addressed at the beginning of this section, one recent theory is that the basilar membrane acts like an acoustic **spectral analyser** (performing a Fourier analysis, see Figure 11.5) or an **acoustic prism**, decomposing the complex signal into the component frequencies by vibrating in specific points, and therefore stimulating the sensory hair cells that sit atop the basilar membrane.

However, measurements on the basilar membrane in living animals found that movements of the basilar membrane are much more localized and much larger than predicted by von Békésy and later experiments have shown that the frequency tuning within the cochlea is not solely due to the passive elastic properties of the basilar membrane. In addition to that, under certain conditions, the ear itself can generate sounds, called *otoacoustic emissions*, which can be detected with a fine microphone placed at the tympanic membrane. Such emissions can occur spontaneously and are believed to be a possible source of *tinnitus* (ringing in the ears). These observations plead for an active biological mechanism which accounts for the ear's sensitivity. The currently available data suggest that the outer hair cell is responsible for an active mechanism of amplification of the acoustic signal. The exact mechanisms of frequency discrimination are still under active research.

2.7. Outer hair cells as cochlear amplifier

The depolarization/hyperpolarization of the outer hair cells in response to the movement of the stereocilia triggers a change in their axial dimensions; depolarization shortens the cells, while hyperpolarization lengthens them (Figure 11.18). This property of outer hair cells is called *electromotility* (electrical to mechanical transduction) and, although the mechanisms is not yet completely understood, it is known to require a protein called *prestin* which is present in high amounts in the membrane of outer hair cells.

Prestin is a special type of biological motor, as its function is not based on enzymatic processes, but on the direct conversion of voltage into displacement (shortening or lengthening), thus its action its much faster than any other molecular motor. Instead of the acoustic signal being dampened by the absorption of its energy by the cochlear structures, the electromotility of

20 For his work, von Békésy received Nobel Prize in Physiology or Medicine in 1961.

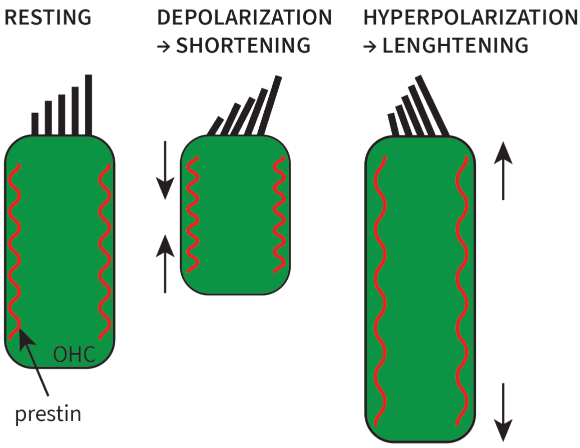


Figure 11.18. The outer hair cells function as a cochlear amplifier. Schematic representation the electromotility of the outer hair cell; prestin molecules sense the changes in membrane potential caused by movement of the stereocilia and convert them into contraction or elongation of the outer hair cell (indicated by the arrows). OHC = outer hair cell.

the outer hair cells amplifies the vibration of the basilar membrane. Overall, the process behaves like adding energy to the acoustic signal, actively amplifying it and narrowing the site of basilar vibration. The functioning of the outer hair cells as a **cochlear amplifier** is vital to everyday communication and for being able to perceive the acoustic richness of speech and music, so for the sharp tuning of the cochlea. Damage of the outer hair cells (induced by large doses of certain antibiotics, for example) significantly reduces the amplification and dulls the sharpness of frequency discrimination.

2.8. Inner hair cells and mechanotransduction of the acoustic signal

The active role of the outer hair cells as a cochlear amplifier facilitates the response of the inner hair cells. Amplification of the basilar membrane vibration augments the shear force between the basilar and the tectorial membrane in specific sites along the basilar membrane, thus stimulating the movement of stereocilia of specific inner hair cells. The **role of the inner hair cells is to convert the acoustic signal (displacement of their stereocilia) into electric signals** which are then relayed via the auditory branch of the vestibulocochlear nerve to the central auditory system.

The mechanically induced change in membrane potential (also called receptor potential) is one of two types: depolarization or hyperpolarization in response to displacement of the stereocilia toward the tallest or shortest stereocilium, respectively. Unlike the outer hair cells, the depolarization of the inner hair cells triggers the opening of voltage gated calcium channels which are present in their basolateral membrane. The increase in intracellular Ca^{2+} concentration stimulates exocytosis of **glutamate** which is stored in vesicles near the basolateral membrane. This excitatory neurotransmitter depolarizes the terminal of the sensory neurons communicating with the inner hair cells and action potentials are triggered and transmitted further towards the brain (Figure 11.19). The greater the amount of neurotransmitter, the higher the rate of action potential firing in the postsynaptic neuron.

The nature of the transduction process in hearing makes the process faster than vision. In hearing, a direct transduction from a mechanical signal (vibration of the hair cells) to an electrical signal is performed, whereas in vision the physical signal (photons) is first converted into a chemical signal (the phototransduction cascade), and finally into an electrical signal. Thus, auditory reaction times are faster than visual reaction times.

The hair cells have the special ability of converting the acoustic signal into electric signal and they are extremely sensitive to vibrations. It has been found that the smallest movement of stereocilia that produces the sensation of sound is ~0.3 nm, which is about the diameter of an atom of gold. In addition to that, the conversion of the displacement into a change in membrane

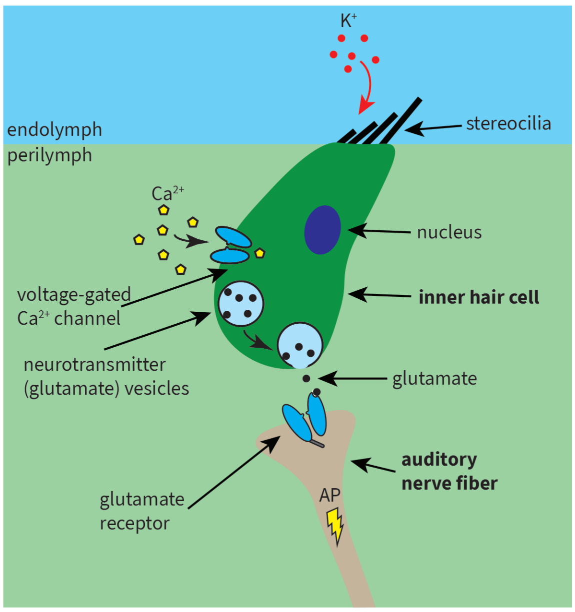


Figure 11.19. Inner hair cells translate the acoustic signal into a receptor potential when depolarized. Lateral movement of the stereocilia opens MET channels, depolarizing the inner hair cell. Depolarization opens voltage-gated Ca^{2+} channels in the hair cell basolateral membrane, allowing entry of Ca^{2+} into the cell. This triggers the fusion of glutamate-containing vesicles with the membrane and their release in the synaptic cleft. Glutamate binds to specific receptors on the membrane of an auditory neuron, and an action potential (AP) can be triggered in the neuron if enough glutamate has been released.

potential is very fast (~10 μs) and this is achieved through direct activation of mechanically gated ion channels present in the tips of the stereocilia. This rapid response allows the membrane potential of the hair cell to change, following the waveform of the acoustic stimulus up to about 3 kHz. Acoustic signals with higher frequencies can still activate the hair cells, although the exact temporal shape of the stimulus is not accurately translated into receptor potentials.

A single inner hair cell can communicate with several auditory nerve fibers, but a single auditory nerve fiber innervates a single inner hair cell. In other words, each auditory neuron sends information about only a tiny part of the basilar membrane vibrating due to the action of a stimulus belonging to a small part of the audible frequency spectrum. Thus, the auditory nerve fibres innervating the base of the cochlea are activated by high frequencies, whereas those leaving from the apical end respond to low frequencies. Electrophysiological recordings performed on such fibres have revealed that each type of fibre responds to a certain range of frequencies and is characterized by a certain **tuning curve**, a graphical representation of the minimum acoustic level required to increase the fiber's firing rate above its spontaneous firing level as a function of frequency. The graph is found to have a characteristic V-shape with a sharp peak corresponding to the minimum threshold value for a given fiber, called the **characteristic frequency**.

In patients with damaged hair cells, the mechano-electrical transduction mechanism is impaired. In such cases, cochlear implants can be used to recreate the patterns of auditory nerve activity elicited by acoustic signals.

REFERENCES

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Berg, J. M., Tymoczko, J. L., & Stryer, L. (2012). *Biochemistry. Seventh Edition*. New York: Freeman and Company.

Boron, W. F., & Boulpaep, E. L. (2017). *Medical Physiology* (3 ed.). Philadelphia: Elsevier.

Ganea, C. (1999). *Elemente de Bioacustică*: Editura Genuine.

Guyton, A. C., & Hall, J. E. (2005). *Textbook of Medical Physiology. Eleventh Edition*. Philadelphia: Elsevier.

Hakizimana, P., & Fridberger, A. (2021). Inner hair cell stereocilia are embedded in the tectorial membrane. *Nature Communications, 12*(1), 2604. doi:10.1038/s41467-021-22870-1

Ivanchenko, M. V., Indzhykulian, A. A., & Corey, D. P. (2021). Electron Microscopy Techniques for Investigating Structure and Composition of Hair-Cell Stereociliary Bundles. *Frontiers in Cell and Developmental Biology, 9*. doi:10.3389/fcell.2021.744248

Parker, A., Parham, K., & Skoe, E. (2022). Noise exposure levels predict blood levels of the inner ear protein prestin. *Scientific Reports, 12*(1), 1154. doi:10.1038/s41598-022-05131-z

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A. S., R.D., M., . . . White, L. E. (2018). *Neuroscience, 6th Edition*. New York: Sinauer Associates.

Suetens, P. (2009). *Fundamentals of Medical Imaging*: Cambridge University Press.

CHAPTER 12

FLUID DYNAMICS AND HEMODYNAMICS

Prerequisite knowledge[1]

▸ Basic notions about the states of matter (solid, liquid, gas)
▸ Pressure. Hydrostatic pressure. Units of measurement (Pa, atm, mmHg)
▸ Density
▸ Friction force, friction coefficient

Matter usually exists in one of three main *states*: solid, liquid or gas. The state in which a substance is found depends on its pressure and temperature. We know from our everyday experience that a solid has a definite volume and shape, a liquid has a definite volume but no definite shape and a gas has no definite volume or shape.

We use the term *fluids* to refer to both gases and liquids. As fluids are composed of molecules held together by weak cohesive forces, they have no particular shape of their own, taking the shape of the container they are in. Additionally, fluids are easily deformable if subjected to an external force, and thus have the ability to *flow*.

The branch of physics that studies fluids is called *fluid mechanics*, which branches into:
▸ **hydrostatics**, which studies fluids at rest;
▸ **hydrodynamics**, which studies fluids in motion.

This chapter will first introduce general notions of hydrodynamics and present the main equations that govern the flow of fluids. In the second part, we will focus on applying these notions to the flow of blood through the circulatory system.

1. HYDRODYNAMICS

1.1. General notions and definitions

There are some specific terms used to characterize the motion of fluids:
▸ *pathline* = the trajectory of an element of the moving fluid;
▸ *streamline* = a curve tangent to the velocity vector, for all the points within the fluid of the same velocity. In a laminar flow (see details later), the pathlines and the streamlines are the same;
▸ *steady flow* = a type of flow in which at any one place in the fluid, the velocity never changes (thus, $dv/dt = 0$). As a consequence, any element passing through a point of specific coordinates has the same trajectory as the previous element that passed through the same point;
▸ *stagnation point* = a point in the flow where the local velocity is zero. This appears, for instance, if flow is blocked by an obstacle.

When a real fluid is in motion, in most of the cases, its flow can be characterized as being one of two main types (Figure 12.1):
▸ **laminar flow**: the adjacent layers of fluid pass one another smoothly, the streamlines are parallel, the paths of different particles never cross each other and for every fluid element arriving at a given point the velocity is the same. The velocity of flow is maximal in the centre of the tube and decreases toward the walls of the tube, where the velocity is 0 (see the section on viscosity for the reason why this happens);
▸ **turbulent flow**: is unsteady, non-permanent, irregular, with whirlpool regions and mass transfer between the streamlines.

The motion of real fluids is very complex and not yet completely understood. For this reason, many times we'll use a simplified version of a fluid, called **ideal fluid**, characterized by:
▸ perfect flow (it has no internal friction, hence no viscosity);
▸ steady flow;
▸ constant density (incompressible);
▸ no angular momentum (irrotational flow).

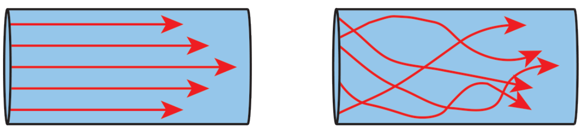


Figure 12.1. Laminar and turbulent flow. Description is provided in the text.

1 Please review these in any introductory Physics textbook, without understanding them this chapter will be extremely difficult to understand. Units of measurement for pressure are also presented in the LP book, in the chapter discussing the measurement of blood pressure.

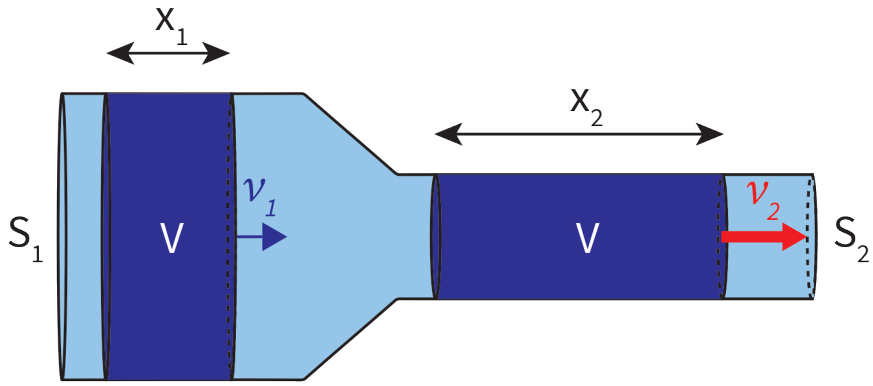


Figure 12.2. The continuity equation. Consider two identical volumes of ideal fluid V that move through the tube through portions of different cross-sectional areas (S_1 and S_2). Thus: $V = S_1 \cdot x_1 = S_2 \cdot x_2$. As volumetric flow is constant, $(S_1 \cdot x_1)/t = (S_2 \cdot x_2)/t$. Writing $x_1/t = v_1$ and $x_2/t = v_2$, we can conclude that $S_1 \cdot v_1 = S_2 \cdot v_2$. Considering that $S_1 > S_2$, we can easily see that $v_1 < v_2$.

1.2. The continuity equation

Let us consider the steady flow of a fluid flowing through a cylindrical tube (Figure 12.2). The volume (*V*) of fluid that flows through the tube in a certain amount of time (*t*) is called the **volumetric flow** (or **flow rate**):[2]

$$Q = \frac{V}{t} \tag{12.1}$$

As the fluid cannot cross the walls of the tube, there is no fluid gained or lost (mass is conserved), so the amount of fluid entering one end must equal the amount of fluid coming out the other end. If the fluid is incompressible, the density of the fluid is constant, thus volume is also conserved (Figure 12.2).

Under these conditions, volumetric flow (*Q*) is constant along the tube and equal to the product of cross-sectional area (*S*) and the velocity of flow (*v*), a relation called the **continuity equation**:

$$Q = S \cdot v = constant \tag{12.2}$$

The continuity equation shows that **when a fluid enters a narrower portion, it will speed up, and it will slow down when it enters a wider portion** (Figure 12.2).

The continuity equation is valid even if the tube has multiple branches. In this situation, the flow velocity depends on the total cross-sectional area of the tube(s) in a certain point and it can be generalized that **the sum of all incoming volumetric flows equals the sum of all outgoing volumetric flows**.

2 Make sure not to mistake volumetric flow (Q) with the flow velocity (v). The volumetric flow is the volume of fluid that flows over the unit of time (measured in m^3/s), while the flow velocity is the distance that the fluid moves on average over time in a particular section of the tube (measured in m/s).

1.3. Bernoulli's equation

Biologists usually say that fluids flow from high to low pressure. This is a simplification which is not always true. When examining fluid flow from the point of view of physics, we have to be more precise. Thus, a more correct statement is that **fluids flow from high potential energy to low potential energy**.

A basic law of physics is **the law of conservation of energy**: energy cannot be created, nor destroyed, but it can be converted into different forms. If this sounds familiar to you, it is because the first law of thermodynamics is essentially a consequence of this more general law.

Let us consider the energy of an ideal fluid flowing through a tube. The total energy of the fluid is the sum of potential and kinetic energies. Thus, for an ideal incompressible fluid, due to the conservation of energy, if the velocity of flow changes, potential energy is converted into kinetic energy and the other way around, but their sum remains constant:

$$E = PE + KE = constant \tag{12.3}$$

where E is the total energy, PE the potential energy and $KE = mv^2/2$ is the kinetic energy.

The rigorous mathematical demonstration of Bernoulli's law is beyond the scope of our text. Let's then consider a simplified explanation. In a fluid at rest (Figure 12.3) the term KE in equation (12.3) is zero, so we consider only the potential energies of the system.

In the example in Figure 12.3, the fluid is at rest. As $KE_1 = KE_2 = KE = 0$:

$$PE_1 = PE_2 = PE = constant \tag{12.4}$$

The potential energy will consist of two terms – a *pressure potential energy* (PPE) and a *gravitational potential energy* (GPE). We know that points which

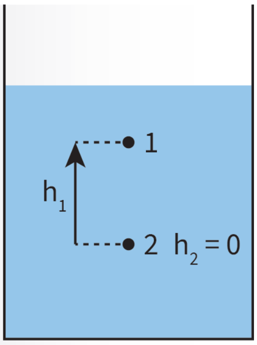


Figure 12.3. A fluid at rest. The law of conservation of energy states that the energy in point 1 has to be equal to the energy in point 2. We arbitrarily chose point 2 as the reference point of the height, so that $h_2 = 0$.

are situated at different heights in a gravitational field have different gravitational potential energies, with $GPE = mgh$. We can then write:

$$PPE + mgh = constant \qquad (12.5)$$

where PPE is the pressure potential energy, m is the mass of the fluid, g the gravitational acceleration and h the height measured from a reference point.

Thus, for the example in Figure 12.3, we can write:

$$PPE_1 + mgh_1 = PPE_2 + mgh_2 \qquad (12.6)$$

To make our life easier, we can take h_2 as the reference point, and, consequently, $h_2 = 0$. Thus:

$$PPE_1 + mgh_1 = PPE_2 \qquad (12.7)$$

It is more convenient to think of fluids in terms of pressure, and not of energy. By dividing the energies by the volume of the fluid (V) we obtain *energy densities*. If we ignore some scientific inconsistencies, we can think of pressure as a type of energy density, as the two have the same units of measurement. We arrive, thus, at:

$$P_1 + \rho g h_1 = P_2 \qquad (12.8)$$

where ρ is the density of the fluid.

Equation (12.8) should look familiar if you previously studied hydrostatics in high school. It essentially says that, the lower you go into a fluid, the higher the pressure will be ($P_2 > P_1$), due to the weight of the fluid (described by the term $\rho g h_1$).

What if the fluid is in motion? We can just add a kinetic energy term. Equation (12.5) becomes:

$$PPE + mgh + \frac{mv^2}{2} = constant \qquad (12.9)$$

We can, again, divide by the volume of the fluid to derive an equation in terms of pressure and arrive at equation (12.10), called Bernoulli's equation.

The conservation of energy during the flow of a fluid is described by **Bernoulli's equation**:

$$p + \rho g h + \rho \frac{v^2}{2} = constant \qquad (12.10)$$

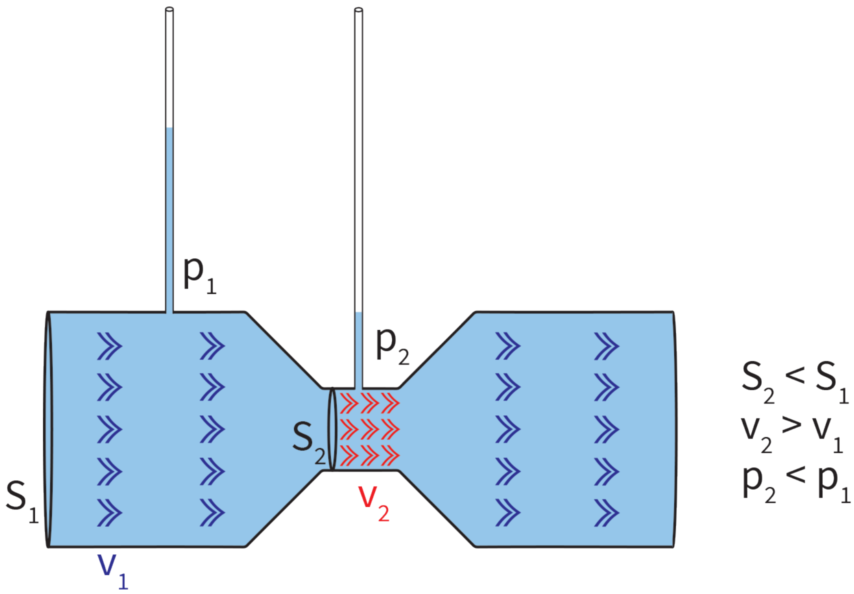


Figure 12.4. The Venturi effect. In narrower parts of the tube, static pressure (p) decreases while the flow velocity (v) increases. This can be visualized by piercing the tube through which the liquid flows with two tubes of much smaller diameter. These tubes are small enough in order to not affect the flow through the main tube. The height of the liquid in the smaller tubes is proportional to the static pressure in the respective sections. Double arrows indicate direction of flow and flow velocity. Dark blue indicates regions where velocity is lower, while red indicates regions where the velocity is higher.

where p is the static pressure, ρgh is the gravitational potential energy density and $\rho v^2/2$ is the dynamic pressure, ρ is density, g is gravitational acceleration, h is the height measured from a reference point and v is flow velocity.

There are three terms in Bernoulli's equation:
- p = ***the static pressure***. We can also simply call this "the pressure", but we add the term "static" here to distinguish it from the dynamic pressure;
- ρgh = the ***gravitational potential energy density***. This only needs to be considered when there is a height variation in the flowing fluid (for example, water flowing downward, like in a waterfall, or water flowing up like in a fountain);
- $\rho v^2/2$ = ***the kinetic energy density***. This is also called the ***dynamic pressure***. See below for a detailed explanation.

In many cases we will consider the flow of fluids placed at a constant height. If $h = 0$, **Bernoulli's equation** becomes:

$$p + \rho\frac{v^2}{2} = constant \tag{12.11}$$

What is **dynamic pressure**? It might be a little counterintuitive, but dynamic pressure is not a real pressure. Rather, it is **a decrease of the static pressure that appears because of the movement of the fluid**. Thus, in the flow of an ideal fluid (without friction), if the flow velocity increases, static pressure will decrease at the expense of increasing dynamic pressure. You can see how that happens when you look back at our demonstration: in order for the kinetic energy of the fluid to increase, its potential energy has to drop.

Using Bernoulli's equation, the **Venturi effect** (Figure 12.4) can be explained: **for a liquid that flows in a tube with a changing diameter, static pressure decreases in the narrow regions of the tube.**

The decrease of static pressure in the narrower portion of the tube in Figure 12.4 occurs because the dynamic pressure increases, due to the increase in the flow velocity. This can happen, for example, in the case of an abnormal narrowing of a blood vessel (*stenosis*). Note that blood flow respects Bernoulli's equation only with a degree of approximation, because blood is a viscous fluid, not an ideal one, while blood vessels are non-rigid tubes (see below). On the other hand, the flow of air in the airways more closely respects Bernoulli's equation.

1.4. Viscosity

The ideal fluid model is useful, but, in many cases, ignoring friction is not a valid assumption. When a real fluid flows, there will always be a certain degree of friction between the wall of the container and the fluid, but also between the adjacent layers of fluid that flow with different velocities. The consequence of the internal friction is the resistance of the fluid to flow, that we call *viscosity*. In other words, viscosity is a measure of the "lack of slipperiness" between layers of fluid, as it was first described by Newton.

The viscosity of a fluid can be determined if one measures the force required to move a plate with constant velocity in that fluid relative to a fixed plate (Figure 12.5A). As an analogy of this experiment, think of the force that you would need to apply in order to make one slice of bread in a creamy peanut butter sandwich slide against the other. This is called the **shear force**.

In Figure 12.5A, we see that the layers of fluid will move at different velocities because of viscosity. There is, thus, a gradient in the velocity (v) across the height of the tube (y) that we call the **shear rate** (noted as $\dot{\gamma}$ and measured in s^{-1}), defined as:

$$shear\ rate = \frac{dv}{dy} \tag{12.12}$$

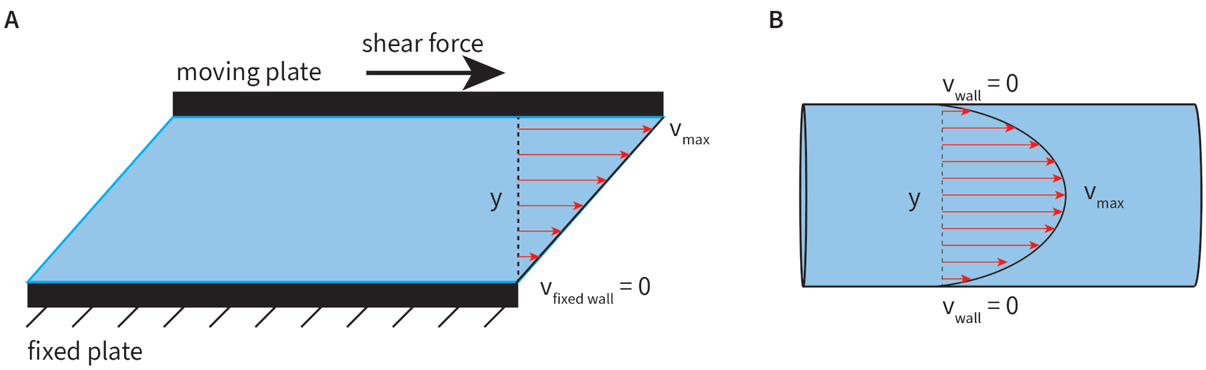


Figure 12.5. Flow of a viscous fluid. A, flow of a viscous fluid held between two plates. No difference of pressure exists initially and the liquid is still. If the top plate is moved, it will cause the liquid to move as well, as the layers of fluid adhere to the two plates and to each other. The layers of fluid will move at different velocities relative to their initial position. B, Flow of a viscous fluid through a tube. The flow profile is parabolic due to the friction between the walls and the fluid layers, as well as friction between the fluid layers themselves.

The *shear force* is proportional to the shear rate and the surface of the moving plate (S) through a quantity that we call the *dynamic viscosity coefficient* or simply **viscosity** (η):

$$shear\ force = \eta S \frac{dv}{dy} \qquad (12.13)$$

In absolute values, the viscosity coefficient (η) is measured in the International System of Units in N·s/m^2, which is equivalent to Pa·s or another unit called poiseuille (PI). However, in practice, another unit is often used, the poise (P), with 1 P = 0.1 PI.

The viscosity of liquids decreases with increasing temperature (for gases, it increases). Thus, the absolute viscosity of water (η_0) at 20 °C is ~0.01 P, while at 37 °C it drops to ~0.007 P. By comparison, the viscosity of blood at 37 °C is ~0.02 P – ~0.04 P.

Another way of expressing viscosity of a fluid is as *relative viscosity*, taking the viscosity of another fluid such as water, as a reference. For example, the relative viscosity coefficient for blood at 37 °C is:

$$\eta_{rel,blood} = \frac{\eta_{0,blood}}{\eta_{0,water}} = \sim 3 - \sim 5.5 \qquad (12.14)$$

where η_0 denotes absolute values of viscosity (measured in Pa·s), while η_{rel} is relative viscosity (dimensionless = has no unit of measurement).

The existence of viscosity will cause most fluids in motion to have a parabolic flow profile (Figure 12.5B). Thus, friction is highest at the boundary between the fluid and the wall, where velocity drops to 0. Conversely, the highest flow velocities will occur towards the middle of the tube. What about shear rates? This might be a little counter-intuitive, but shear rates are highest at the wall and are lowest in the middle of the tube. This is because the shear rate shows the change in velocity over the width of the tube, not the velocity itself (Figure 12.5B).

Equation (12.13) holds true only for fluids that have a constant viscosity, regardless of the shear rate (e.g. water, air). There are also fluids for which the viscosity is not constant, but changes if the shear rate changes and there is no direct proportionality between shear force and shear rate (e.g. paint, ketchup, cornstarch and water mixture). Thus, fluids are classified as:

- **newtonian**: their viscosity does not change with shear rate;
- **non-newtonian**: their viscosity changes with the shear rate.

It may be easier to understand the difference if we compare the flow of two fluids from different categories: water and ketchup. Take a bottle of water, remove the lid and turn it upside-down over a sink. Water will instantly start to flow. Now

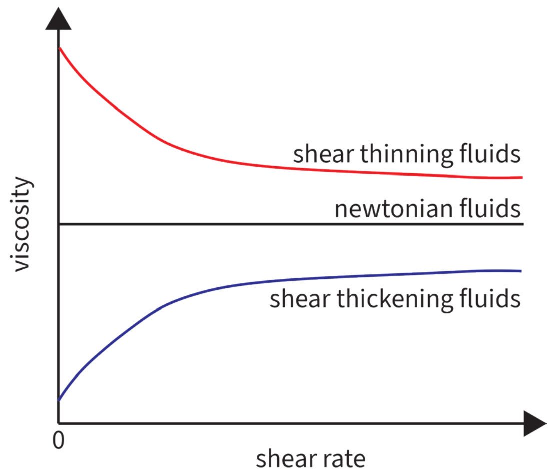


Figure 12.6. Viscosity *vs* shear rate for different types of fluids.

do the same experiment, but shake the water bottle beforehand. When you turn the shaken bottle upside-down, nothing has changed, and water will flow in a similar manner as in the first example. Ketchup behaves quite differently from water. Turn an unshaken bottle of ketchup upside-down and very little (if any) ketchup flow will occur. Shake the bottle, however, and ketchup will pour out a lot easier. In fact, it might spill all over your clothes, so be careful!

We can, thus, conclude that the flow behavior of these two fluids is different: water has the same viscosity at different shear rates (it's a newtonian fluid), while the viscosity of ketchup decreases at higher shear rates (it's a non-newtonian fluid). Depending on the relationship between viscosity and shear rate (Figure 12.6), non-newtonian fluids can be:

- *shear thickening* fluids: viscosity increases with increasing shear rate , e.g. a mixture of cornstarch and water;
- *Bingham plastic* fluids: they require a threshold force to get them flowing, e.g. mayonnaise;
- *shear thinning* fluids (also called *pseudoplastic* fluids): viscosity decreases with increasing shear rate, e.g. ketchup, blood, synovial fluid.

The **synovial fluid** that fills the synovial joints is another example of a non-newtonian shear thinning fluid. The presence of a fluid between two moving solid surfaces greatly attenuates friction between them; this effect is called *lubrication*. The synovial fluid reduces friction between cartilage surfaces, thus it acts as a lubricant.

Due to its composition (lubricin, hyaluronic acid, etc.), the synovial fluid lubricates joints over a wide range of muscle force and velocity, by adapting its viscosity to the type of motility the joint is performing. When a muscle attached to a joint is bearing a light load, the joint movement can be rapid (high shear rates) and, in order for the synovial fluid to flow easily to lubricate all the

surfaces, its viscosity must be low. In contrast, when a muscle is bearing a large load, the joint movement will be slow (low shear rates) and the viscosity of the synovial fluid will considerably increase, reaching a gel-like state. This increase in viscosity prevents damage to the articular surfaces and the flow of synovial fluid away from the joint when large loads are borne.

1.5. Poiseuille's law

The flow of a viscous fluid along a tube (a pipe, a blood vessel, a hypodermic needle, etc.) requires a pressure difference at the ends of the tube to overcome the resistance of the fluid to flow. For example, if you drink a certain volume of thick smoothie using a straw, your effort will be greater than if you drink the same volume of water in a certain period of time. A smoothie is more viscous than water, so it has a higher resistance to flow, hence a stronger pressure difference is needed to get it flowing. Also, if you pick a narrower straw, it would be more difficult for you to sip the same volume of smoothie in the same period of time as before. The same will happen if you pick a longer straw, due to an increase in the solid-fluid friction surface. You might then conclude that, for a constant volumetric flow along a tube, the necessary pressure difference will be the higher if viscosity of the fluid increases, the radius of the tube decreases or the length the tube increases.

All these parameters are quantitatively related by **Poiseuille's Law**, which states that the volumetric flow (Q) for a fluid of viscosity η along a cylindrical rigid tube of length l and radius r, with a difference in pressure ΔP between the ends, is:

$$Q = \frac{\pi r^4 \Delta P}{8 \eta l} \quad (12.15)$$

It is important to notice that the volumetric flow is directly proportional to the radius to the fourth power which means that a small change in the radius would cause a significant change in the volumetric flow. We will return to this law later, when we talk about blood flow.

1.6. Stokes' law

A change in the viscosity of blood can be indicated by a change in the velocity at which the red blood cells fall down in a blood-filled tube, a test routinely performed in the clinic[3]. The relation between the viscosity of blood and the velocity of the falling red blood cells is described by an equation called **Stokes' law**:

$$F = 6\pi\eta r v \quad (12.16)$$

Stokes' law states that if a sphere with the radius r is moving with velocity v in a fluid with viscosity η, then the fluid will exert a force F to oppose the falling of the sphere. This force is called the *frictional force* or the *drag force*.

1.7. Reynolds number

As already described in the beginning of this chapter, there are two main types of flow for a viscous fluid flowing through a tube, *laminar flow* and *turbulent flow* (Figure 12.1). The way in which a viscous fluid flows through a tube depends on its viscosity (η), density (ρ) and velocity of flow (v), as well as on the diameter (d) of the tube. We can define a dimensionless number called the **Reynolds number (Re)**, which shows whether flow is laminar or turbulent:

$$Re = \frac{\rho v d}{\eta} \quad (12.17)$$

Thus:

- at low Reynolds numbers (< ~2000), flow is laminar;
- for values of Re = ~2000 – ~3000 we call the flow transient: it is a transition stage between a laminar flow and a turbulent flow that has characteristics of both;
- if the Reynolds number increases above ~3000, flow becomes turbulent.

2. HEMODYNAMICS

Hemodynamics is the study of the physical laws of blood circulation. It takes into consideration the properties of both the fluid (blood) and the container (blood vessels) and it describes the relationship between blood pressure, blood flow and the resistance to blood flow.

2.1. The circulatory system

The circulatory system is a closed tubular system in which the heart acts as a pump, generating the pressure gradient required to get blood flowing through blood vessels with variable diameter and elastic walls.

The general schematic of the circulatory system (Figure 12.7) shows a large number of vessels through which blood flows from the region with high blood pressure (aorta) toward the region with low blood pressure (the two venae cavae). The volumetric blood flow is on average equal to 5 L/min at rest and it can increase up to 25 L/min

3 The test is called the *erythrocyte sedimentation rate* (ESR).

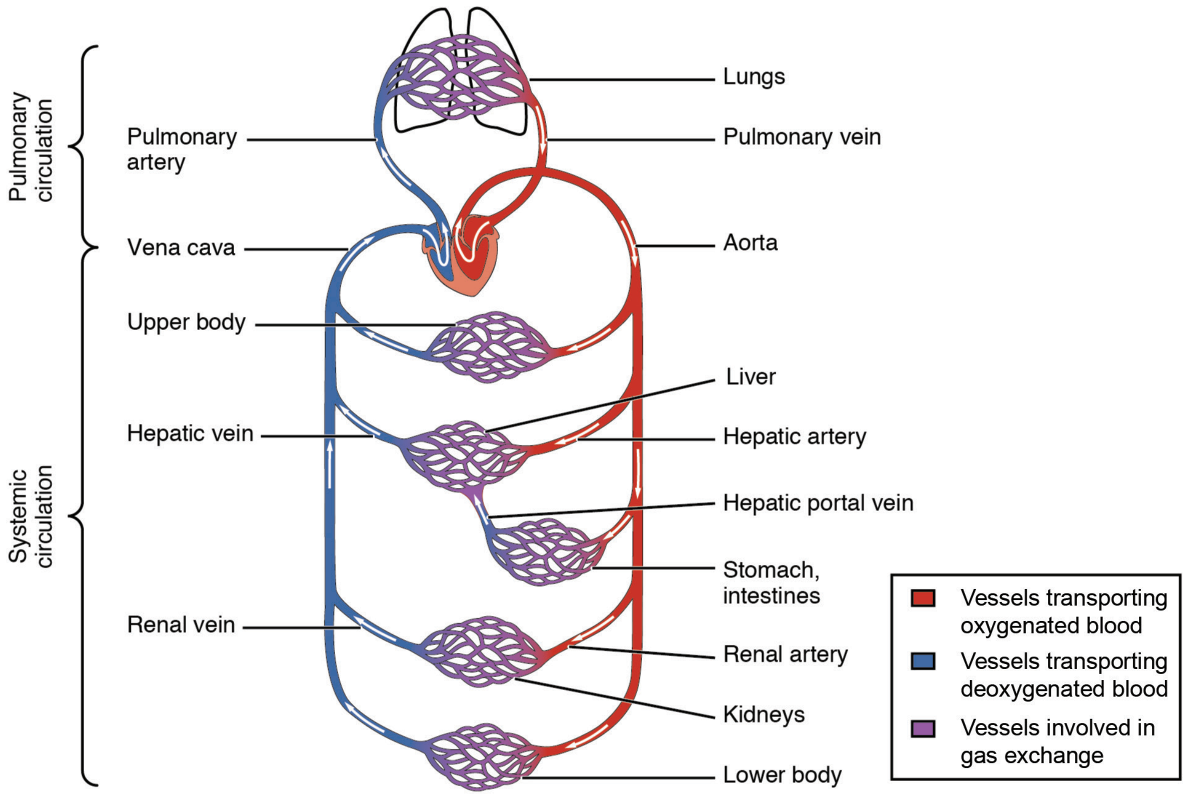


Figure 12.7. The circulatory system.[4]

during physical effort. The aorta branches into arteries and the small arteries into arterioles. From arterioles, blood flows through the capillary bed toward venules, veins and finally returns to the heart via the venae cavae.

As it leaves the heart, blood flows through smaller and smaller blood vessels. Although the diameter of individual blood vessels decreases, the number of branches increases, so the overall cross-sectional area also increases. The total capillary cross-sectional area is ~1000 times larger than that of the aorta (Table 12.1).

The volumetric blood flow in any point of the circulatory system is constant and depends on the pressure difference between the average aorta pressure (100 mmHg) and the average pressure in the venae cavae (10 mmHg).

An important site for controlling blood flow is the **pre-capillary sphincter**, a ring-shaped smooth muscle that, through its contraction and relaxation, regulates the diameter of the blood vessel and thus the blood flow from arterioles to capillaries (Figure 12.8). This muscle is sensitive to changes in the conditions the surrounding tissue. For example, a decrease in O_2 concentration, decrease in pH, increase in CO_2 or in K^+ concentrations determine the muscle to relax, allowing the flow of blood from the arteriole into the capillaries within the tissue. When opposite conditions are met, the muscle contracts and the flow is reduced. Thus, blood does not flow smoothly through a capillary, but in an irregular, pulsating manner.

2.2. The structure of the vascular walls

With the exception of the capillaries, the walls of the blood vessels are composed of three layers of tissues (tunics, Figure 12.9) named, from the

Table 12.1. Total cross-sectional areas[5] for different types of blood vessels.

Type of blood vessel	Cross-sectional area (cm^2)
Aorta	2.5
Small arteries	20
Arterioles	40
Capillaries	2500
Venules	250
Small veins	80
Venae cavae	8

4 Modified from an image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). Anatomy and Physiology 2e. Retrieved from *https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction*

5 According to Guyton, A. C., & Hall, J. E. (2005). Textbook of Medical Physiology. Eleventh Edition. Philadelphia: Elsevier.

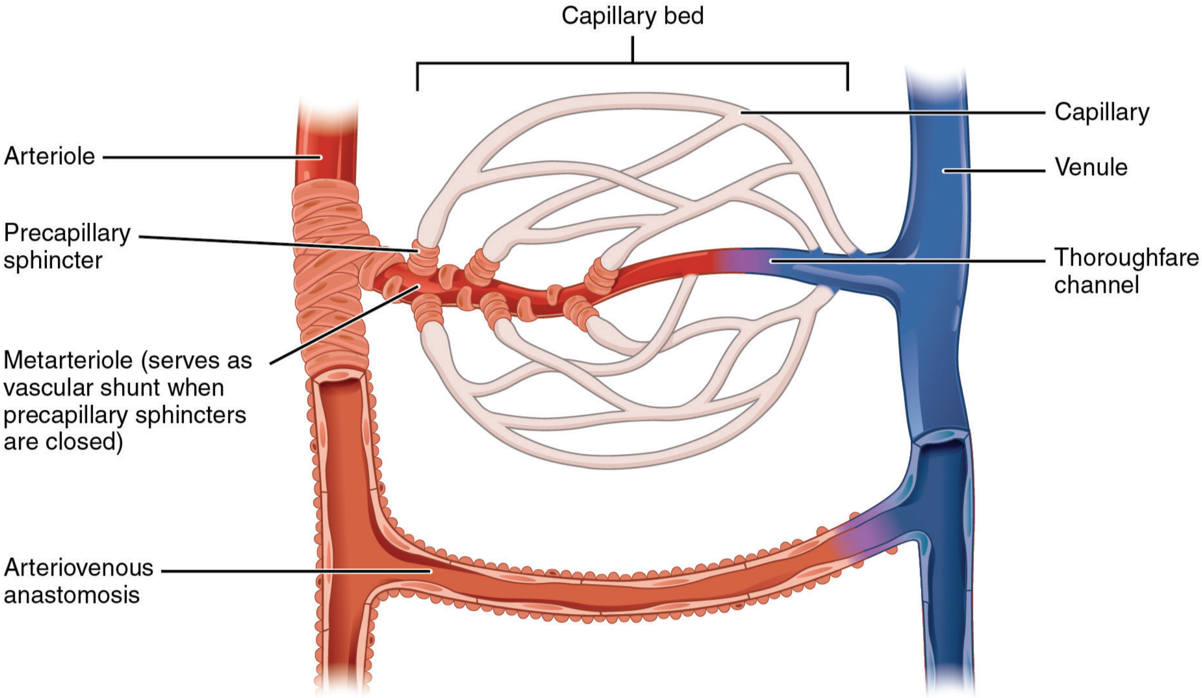


Figure 12.8. Pre-capillary sphincters.[6]

lumen towards exterior: the *intima*, the *media*, and *tunica externa* (the *adventitia*), with different composition for different blood vessels. Capillaries have only an intimal layer of endothelial cells resting on a basal lamina.

Regardless of the particularities of the tunics, there are four building blocks that make up the vascular wall:

- **The endothelial cells** form a single, continuous layer (the intima) that lines the lumen of the blood vessel, providing a smooth surface and also a selective permeability for various substances (water, electrolytes, glucose, etc.);
- **Elastin fibers** are part of the media. They are easily extensible and contribute to the *passive tension* that arises in the vascular wall due to the static pressure exerted by blood flowing through the vessel. Elastin fibers are present in all types of blood vessels except for capillaries, venules and arteriovenous anastomoses;
- **Collagen fibers** are present in the media as well as in the adventitia. They are more rigid than elastin fibers, so they are important for the resistance to stretch of the vascular wall at high blood pressures, thus contributing to the passive wall tension. Collagen fibers are of great importance in arteries;
- **Smooth muscle fibers** are part of the media. Except for the capillaries, they are present in all vascular segments. However, their content is maximal in arterioles where they form the pre-capillary sphincters. Due to their ability to contract and relax depending on the state of the surrounding tissue, the smooth muscle fibers are responsible for the *active tension* of the vascular wall.

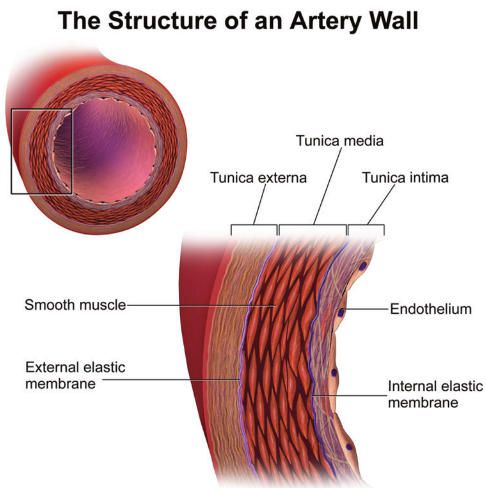


Figure 12.9. The tunics of an arterial wall.[7]

6 Image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). Anatomy and Physiology 2e. Retrieved from *https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction*

7 Copyright-free image provided by Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". WikiJournal of Medicine 1 (2). DOI:*10.15347/wjm/2014.010*. ISSN 2002-4436.

2.3. Laplace's law

When blood flows through blood vessels, it exerts a certain pressure onto the vascular walls (the static pressure from Bernoulli's equation). There is also an external pressure which opposes the pressure inside the blood vessel. The net pressure onto the vascular wall is the difference between these two types of pressure and is called transmural pressure, Δp. Because the vascular walls have elastic components, they are non-rigid and a vascular wall tension, T, appears. This is a force[8] in the wall of the blood vessel that resists the stretching (distension) of the walls due to transmural pressure. The equilibrium between Δp and T depends on the blood vessel radius, r, and can be mathematically expressed for a cylindrical tube by **Laplace's law**:

$$\Delta p = \frac{T}{r} \tag{12.18}$$

The above equation is valid for a cylindrical tube with negligible thickness relative to the tube's radius. Because blood pressure is much greater than the external pressure, we can approximate transmural pressure with blood pressure. From this point of view, Laplace's law describes the tension that arises in the vascular wall for a given value of the blood pressure. In other words, changes in the blood pressure can cause changes in the wall tension.

For elastic membranes with variable curvatures, the law of Laplace takes into account the two main radii (r_1 and r_2):

$$\Delta p = T\left(\frac{1}{r_1} + \frac{1}{r_2}\right) \tag{12.19}$$

For spherical structures with elastic walls, because $r_1 = r_2 = r$, Laplace's law becomes:

$$\Delta p = \frac{2T}{r} \tag{12.20}$$

This equation should look familiar to you, because it was discussed in a previous chapter in order to explain the relation between the alveolar surface tension (the equivalent of T in the above relation), the alveolar pressure and the radius of the alveolus.

Although the heart is not spherical but rather conical, Laplace's law can be useful, with a certain degree of approximation, to describe the relation between ventricular pressure and wall tension. The left ventricular wall is thicker than the wall of the right ventricle. This enables the left ventricle to develop a stronger tension, but also to reduce the dimension of the ventricular cavity through contraction, thus greatly increasing the ejection pressure. The left ventricle has a conical shape, with a much smaller radius of curvature than the right ventricle. This makes it easier for the left ventricle to generate high pressures, as a lower tension has to be developed in the ventricular wall, according to the law of Laplace.

2.4. Distensibility and compliance

Blood vessels are *distensible*: when the transmural pressure increases, blood vessels dilate, lowering their resistance to flow. We can calculate the **distensibility** of a blood vessel as:

$$distensibility = \frac{\Delta V}{V_0 \cdot \Delta p} \tag{12.21}$$

where ΔV is the change in volume, V_0 is the initial volume and Δp is the change in pressure.

Thus, distensibility can be thought of as the degree to which a vessel dilates when subjected to an increase in blood pressure. For example, a distensibility of 0.1 $mmHg^{-1}$ corresponds to an increase of the volume by 0.1 of the initial volume (10%) when blood pressure increases by 1 mmHg.

In most cases, it is more convenient to calculate a related quantity called **compliance**, which shows not the relative, but the absolute increase in volume with pressure. This is calculated as:

$$compliance = distensibility \cdot V_0 = \frac{\Delta V}{\Delta p} \tag{12.22}$$

where ΔV is the change in volume and Δp is the change in pressure.

A remarkable difference exists between the compliances of veins and arteries: **veins are about ~24 times more compliant than their corresponding arteries**. This is explained by the thicker and stronger walls of the arteries, which do not allow them to expand as much (lower distensibility) as well as by the larger volume of veins compared to that of arteries. This makes veins very well adapted to serve as a reservoir of blood inside the body.

2.5. Blood pressure (BP)

During ventricular systole, the pressure in the left ventricle is very high and, when blood is ejected into the aorta, both an *arterial pressure wave* (which propagates through the arterial wall) and a *blood flow wave* are generated. The pressure wave is generated by the distension of the aortic walls during systole and it propagates down the

8 Tension (T) is often defined as the force (F) that needs to be applied to bring together the two edges of an imaginary slit of unit length L, cut along the longitudinal axis of the vessel. Thus, T = F/L. Do not confuse it with pressure, p = F/S, where S is the surface area.

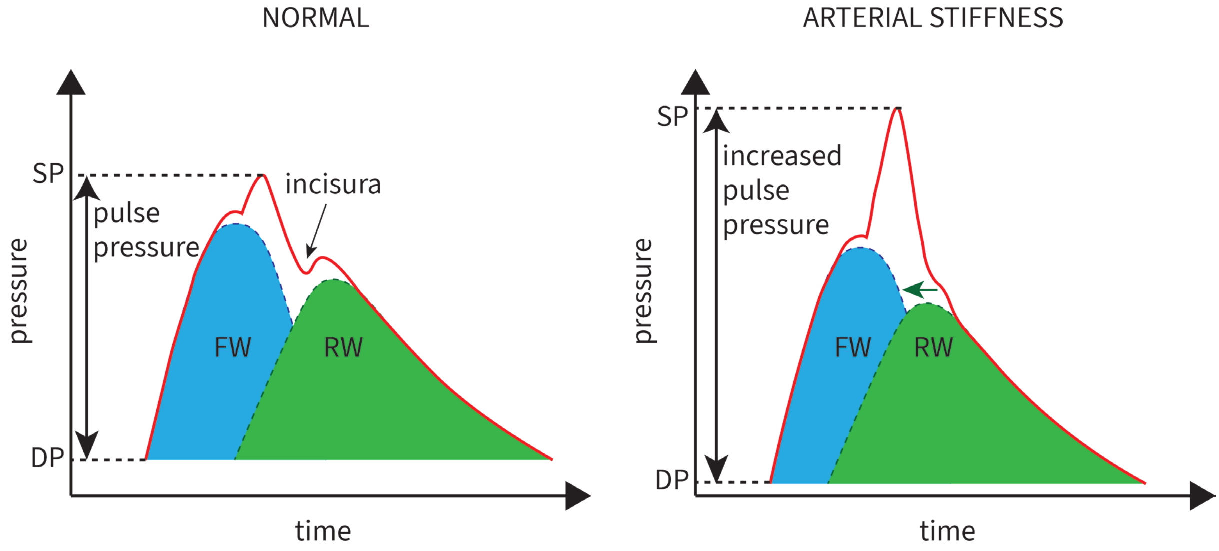


Figure 12.10. Pulse pressure wave in the aorta during a cardiac cycle. Left: normal pressure waveform. Right: pressure waveform under conditions of increased arterial stiffness. The forward (FW) and reflected (RW) waves add up to form the aortic pressure (continuous red line). The green arrow in the right panel indicates that, when arteries stiffen, the reflected wave arrives earlier, and is summed up with the transmitted wave. SP – systolic pressure, DP – diastolic pressure.

arteries much faster than the mean blood velocity. When you press your radial artery, it is the arterial pressure wave that you feel as pulse and not the blood flow wave. If this is hard to grasp, just think what travels faster: an ambulance that is moving towards you or the sound (acoustic wave) from its sirens?

The aortic walls distend and store part of the ejection volume of blood, raising the arterial pressure up to the systolic ventricular pressure. During ventricular diastole, the previously stretched aortic walls recoil passively and the stored volume of blood is driven towards capillaries. The aortic volume of blood, and, consequently, the aortic pressure, both decrease slowly as blood flows away, but they never fall to zero because the next ventricular systole occurs while there is still blood in the aorta. These periodical changes in blood pressure (pulsatile pressure), are not only characteristic of the aorta but of all large arteries, because the pulsatile pressure travels as a pressure wave down the arterial tree. The maximum value of arterial blood pressure, the **systolic pressure** (SP), is reached during the peak of ventricular systole while the minimum, the **diastolic pressure** (DP) occurs at the end of ventricular diastole. Despite the pulsatile pressure, due to the elastic properties of the arterial walls, tissues and organs are perfused at all times.

Figure 12.10 shows the aortic pressure variation during a cardiac cycle. The ascending phase starts when the aortic valve opens and blood is ejected into the aorta. The *forward pressure wave* (also called *transmitted wave* or *incident wave*) travels through the aortic wall and, when the wavefront encounters a change in the vessel resistance, a *reflected wave* is generated and travels back, towards the aortic valve. The resistance to flow of the large arteries is relatively constant, but increases significantly at the site of the arterioles due to their high content in smooth muscle. Thus, the arterioles are the major site for the reflection of the pressure wave.

The pulse pressure at any point in the walls of the arterial system is given by the summation of the forward and the reflected waves (continuous red line in Figure 12.10). Its amplitude and shape depend on the speed at which the pressure wave travels and the strength of the reflected wave. The maximum of the forward wave corresponds to a first inflection point, followed by an increase in pressure (due to the addition of the reflected wave) up to a maximal value, the **systolic pressure**. At the end of the ventricular systole, before the closure of the aortic valve, the aortic pressure starts declining due to the attenuation of the transmitted wave. When the aortic valve closes, the reflected wave reaches its peak, which contributes to a transient increase (called the *incisura*) in the total pressure. After that, the aortic pressure continues to fall and reaches a minimum value, the **diastolic pressure**, just before the next opening of the aortic valve. The diastolic pressure is the pressure exerted by the vascular tree onto the aortic valve.

Any changes in the cardiac activity or the vascular properties can alter the shape of the pressure wave. For example, the slope of the initial ascending part decreases in case of suboptimal myocardial contraction or of aortic stenosis. The slope of the final descending part (at the end of

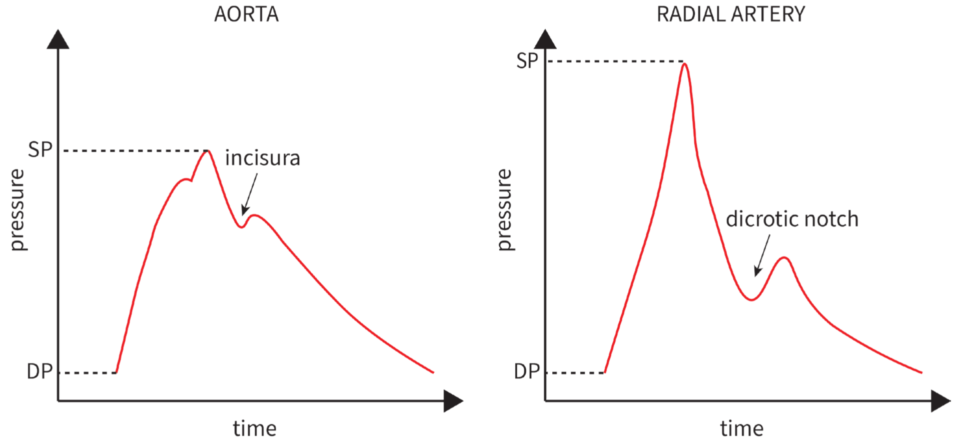


Figure 12.11. Blood pressure waveforms in different arteries. Left: pressure waveform in the aorta. Right: pressure waveform in the radial artery. The systolic pressure is higher in the radial artery. Also, the incisura disappears and a similar shape appears on the waveform called the dicrotic notch, which results more from the resistance of peripheral vessels rather than the closing of the aortic valve. SP – systolic pressure, DP – diastolic pressure.

diastole) also decreases in case of aortic stenosis. The elasticity of the arterial walls decreases with increasing age, decreasing compliance and leading to a stronger reflection. This causes an increase in the systolic pressure, a decrease in the diastolic pressure and the attenuation of the incisura by comparison to physiological conditions (Figure 12.10, right panel). Overall, this results in an increased *pulse pressure*, which is defined as:

$$pulse\ pressure = SP - DP \qquad (12.23)$$

In addition to that, the recoil reaction of the arterial walls after being stretched happens faster, if the vascular wall is more rigid. Therefore, the reflected wave reaches the aortic valve sooner, during systole and not during diastole like in the case of vascular walls with normal elasticity, thus increasing the ventricular effort and decreasing the ability to transport oxygen and nutrients through the coronary arteries.[9]

In general, the elasticity of the vascular wall decreases with distance from the aorta. As a consequence of that, the pressure variations due to reflections of the pressure wave are enhanced, thus the amplitude of the pressure wave increases,

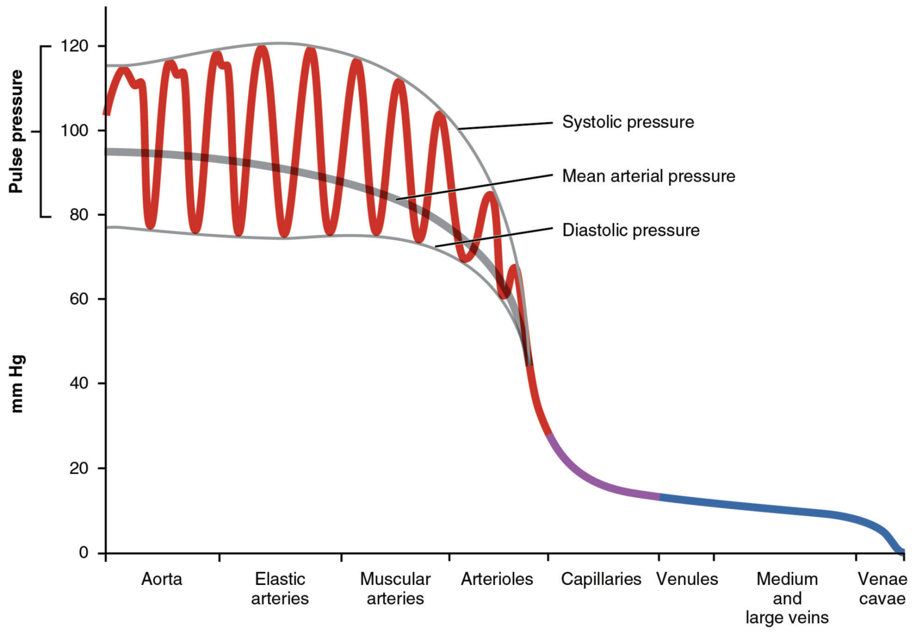


Figure 12.12. Blood pressure variation in different types of blood vessels.[10]

9 Coronary arteries transport blood to the heart muscle.
10 Image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). Anatomy and Physiology 2e. Retrieved from *https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction*

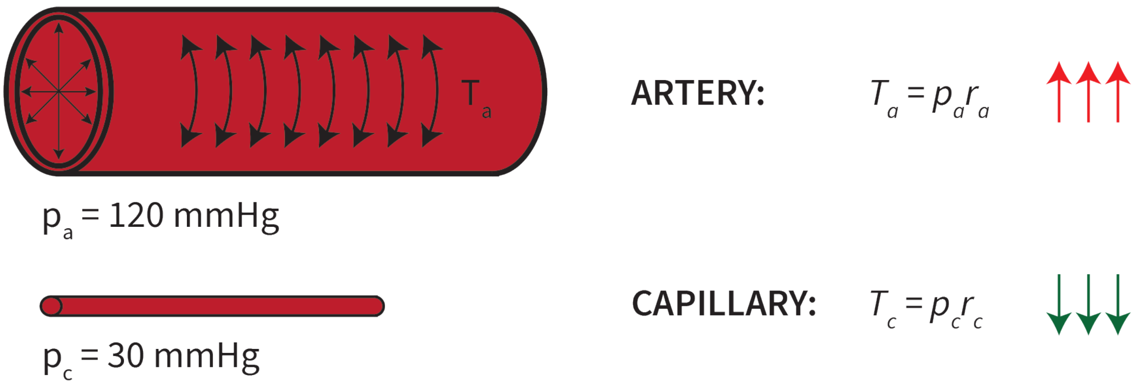


Figure 12.13. Wall tension in arteries and capillaries. According to the law of Laplace for a cylindrical blood vessel, T = p·r. In capillaries, due to the much smaller radii, as well as the lower blood pressures, wall tension is much lower than in arteries. T_a = arterial wall tension, T_c = capillary wall tension, p_a = arterial pressure, p_c = capillary pressure, r_a = artery radius, r_c = capillary radius.

causing an increase of the systolic pressure in the peripheral vascular areas (Figure 12.11). Relative to the aorta, these variations can have values of 20 – 30 mmHg at the brachial artery or 40 mmHg at the radial artery. The diastolic pressure doesn't change significantly from the aorta to the radial artery because the vascular wall elasticity doesn't decrease much. On the other hand, on the path of ascending/thoracic/abdominal aorta – iliac artery – femoral artery, the elasticity of the vascular wall decreases significantly and, because of that, the systolic pressure increases, while the diastolic pressure progressively decreases.

Regardless of the systolic/diastolic pressure changes, due to the increase in the resistance to flow, the average pressure drops progressively from the aorta towards arteries, then arterioles and capillaries. The greater the change in resistance at any point in the vasculature, the greater the loss of pressure at that point. Arterioles have the greatest increase in resistance and cause the largest decrease in blood pressure (Figure 12.12). In veins, blood pressure is quite low, because most of pressure imparted by the ventricle contraction was dissipated by the resistance and it continues to fall until it reaches the venae cavae.

The average arterial pressure throughout one cardiac cycle is called the **mean arterial pressure** (MAP). Strictly speaking, this can be calculated from the area below the pressure curve. However, in a commonly used approximation, this is calculated from the values of the SP and DP, taking into account the fact that, in a cardiac cycle, the duration of diastole is approximately double relative to systole:

$$MAP = \frac{SP + 2DP}{3} \qquad (12.24)$$

The MAP is essential for the perfusion of organs in the body. Normal values are of ≥ ~70 mmHg. At MAP <60 mmHg organs will not receive enough blood, leading to hypotensive shock and organ failure. Some representative values for the mean pressures in various blood vessels are: 90 – 100 mmHg in the aorta, 30 – 40 mmHg in arterioles, 15 – 20 mmHg in capillaries, 10 – 15 mmHg in venules and 3 – 8 mmHg in the venae cavae.

Arterial pressure can be measured directly (an invasive method) or indirectly, through non-invasive methods. The direct method involves the insertion into the artery of a narrow tube (catheter) which has a miniaturized manometer[11] and it is more rarely used (for example, during reanimation procedures). Indirect methods include: the palpatory method, the auscultatory method and the oscillometric method.

The measured blood pressure values are reported as SP/DP mmHg. For example: 120/80 mmHg, 106/66 mmHg, etc.

As stated above, the values measured for the arterial pressure depend on the site of measurement. In general, the pressure measured at the brachial artery is taken as reference. The radial artery is another possibility for an easy measurement. However, the pressure values measured there have a large variability. The range of values which are currently considered normal by the European Society of Cardiology for a healthy adult <65 years old are: **90 – 140 mmHg** for the systolic pressure and **60 – 90 mmHg** for the diastolic pressure, both measured at the level of brachial artery. The normal upper limit value for the central (at the level of the aorta) systolic pressure increases with age: 110 mmHg for the age of 20 – 30 years and 125 mmHg for the age of 50 – 60 years.

If either the SP or DP are higher than normal, the patient suffers from **hypertension**. If either the SP or DP are lower than normal, the patient suffers from **hypotension**.

Remember that we talked above about Laplace's law in equation (12.18), which corelates

11 A manometer is a device that measures the pressure of a fluid.

the pressure of a fluid flowing through a cylindrical tube with the tension in the wall of the tube and its radius. According to this law, the tension in the arterial walls is very large, while very small tensions appear in the capillary walls (Figure 12.13). For a brachial systolic pressure of 120 mmHg, the blood pressure at the entry into the capillary network is relatively high (around 30 mmHg), but due to the small radii of the capillaries, their wall tension is small. Because arteries need to withstand large pressures, their walls have a high content in collagen fibers, ensuring a high resistance to stretching and thus protecting them from aneurysm (see below).

2.6. Factors that influence arterial blood pressure

- **Heart rate**: the higher the heart rate, the higher the blood pressure.
- **Total blood volume**: the higher the total blood volume, the higher the heart rate, so the higher the blood pressure.
- **Heart volumetric flow** (also called *cardiac output*). This is the product of heart rate and the *stroke volume* (the volume of blood delivered by the heart during each contraction). The arterial blood pressure increases with heart volumetric flow. For example, at rest, for a cardiac output of 5 L/min, the arterial blood pressure is around 125/80 mmHg, while during physical effort, if the cardiac output increases by 15 L/min, blood pressure can have values of 180/125 mmHg.
- **Resistance to flow**. If you went back to the imaginary experiment described in section 1.5, about the pressure difference at the ends of the straw required to overcome the resistance of the fluid to flow, you would conclude that, for a constant volumetric flow, if resistance to flow (R) increases, the pressure difference must increase accordingly:

$$R = \frac{\Delta p}{Q} \qquad (12.25)$$

From the same experiment we saw that the resistance to flow is given by the viscosity of the fluid but also by the dimensions of the tube. Considering the equation above and Poiseuille's law in equation (12.15), we can mathematically describe resistance to flow as:

$$R = \frac{8\eta l}{\pi r^4} \qquad (12.26)$$

Blood vessels are not rigid tubes and the relationship between the radius and the volumetric flow is more complex. If, for example, the blood volumetric flow in a normal artery was 100 cm^3/min at a pressure of 120 mmHg, then a reduction by 20% of the arterial radius (e.g., due to an atheroma plaque) would cause a decrease in the volumetric flow to 41 cm^3/min, and the pressure required to reestablish the normal volumetric flow would be 293 mmHg.

To give another example, if during performing physical effort the blood volumetric flow increased 5 times and vasodilation processes weren't possible, blood pressure should increase from 120 mmHg to 600 mmHg in order to maintain this flow. However, the increase in volumetric flow is possible without any change in blood pressure, due to vasodilation of the arterioles. Thus, an increase of 1.5 times in the arteriolar radius would lead to an increase in blood volumetric flow of $(1.5)^4$, so about 5 times. The arteriolar system has a great contribution in controlling the blood volumetric flow (Figure 12.14). These small, but numerous blood vessels, can reduce or increase blood flow toward some parts of the body (through vasoconstriction and vasodilation, respectively), according to tissular needs.

Under normal physiological conditions, blood

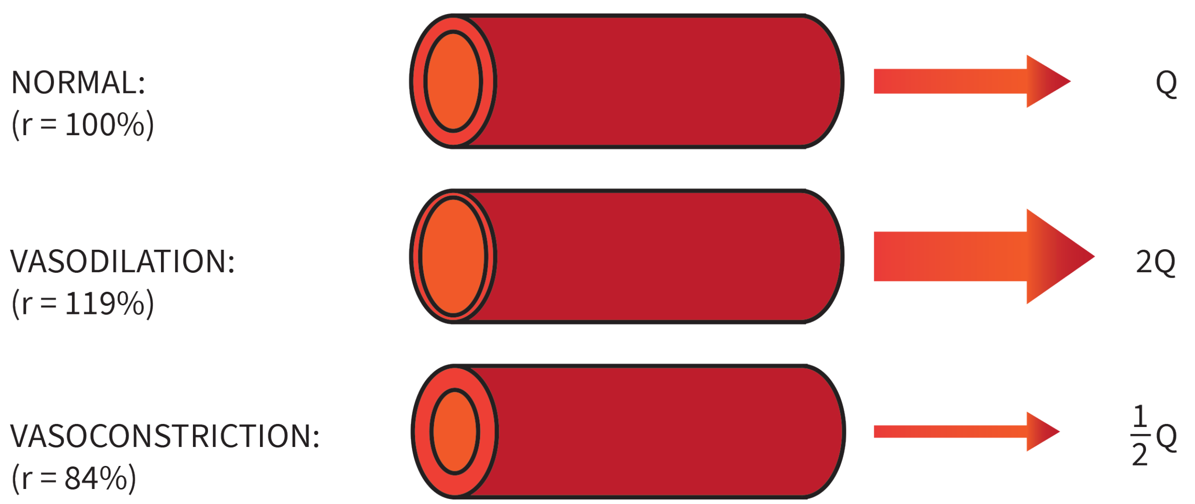


Figure 12.14. Effect of blood vessel radius on blood flow. A 19% increase of the radius doubles blood flow. A 16% decrease of the radius reduces blood flow to half.

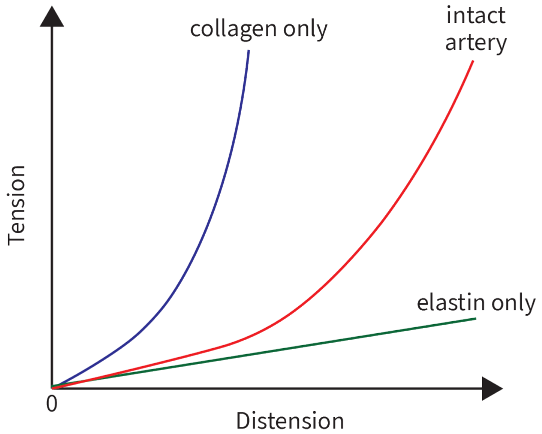


Figure 12.15. Role of elastin and collagen in the distensibility of arteries. The "collagen only" curve represents the arterial wall treated with trypsin (dissolves elastin). The "elastin only" curve represents the arterial wall treated with formic acid (dissolves collagen).

volumetric flow is constant. Thus, an increase in the resistance to flow (for example, due to a reduction in the diameter of the blood vessel, or due to a less smooth lumen surface) leads to an increase in Δp.

▸ **Elasticity of arterial walls**: the arterial pressure increases if the rigidity of the arterial walls increases. There is a large category of antihypertensive drugs which reduce the blood pressure by decreasing the arteries' rigidity and preventing vasoconstriction. These drugs block the calcium channels from the plasmalemma of the smooth muscle fibres present in the vascular wall, leading to the relaxation of the fibres.

Due to the complex structure of the arterial wall, its elasticity is not constant, but decreases with increasing arterial blood pressure, so the relationship between arterial wall tension and distension[12] is not a linear one (Figure 12.15). At small pressures, arteries dilate easier, but at larger pressures the vascular walls become more resistant to stretch. The elasticity of arterial walls decreases with age and, for this reason, blood pressure increases with age.

▸ **Blood viscosity**: if blood is more viscous, then the arterial pressure is higher.

Blood is a complex liquid, a suspension of cells in plasma (Figure 12.16). The relative viscosity of isolated plasma (a newtonian fluid) is only ~1.8 whereas blood is 3.5 – 5 times more viscous than water and behaves as a **non-newtonian shear thinning (pseudoplastic)** fluid. Its viscosity depends primarily on *hematocrit, H* (the percentage by volume of red blood cells), but also on the diameter of the blood vessel and the velocity of flow. Blood viscosity increases almost exponentially with the hematocrit (Figure 12.17). A too low hematocrit value would cause a decrease in blood viscosity, while a value higher than normal is associated to an increase in the viscosity of blood. The *optimal value of hematocrit* (~48% for men, ~41% for women) corresponds to the highest amount of hemoglobin in the capillaries and it is proportional to the maximum H/η ratio.

The **shear thinning properties of blood** are given by the flow behavior of erythrocytes. At low shear rates, red blood cells form aggregates that make viscosity high. These are progressively broken down as shear rate increases. At high shear rates, red blood cells tend to orient themselves along the flow direction *(axial accumulation)* and can even become deformed (elongated) at very high shear rates. Because of that, the resistance

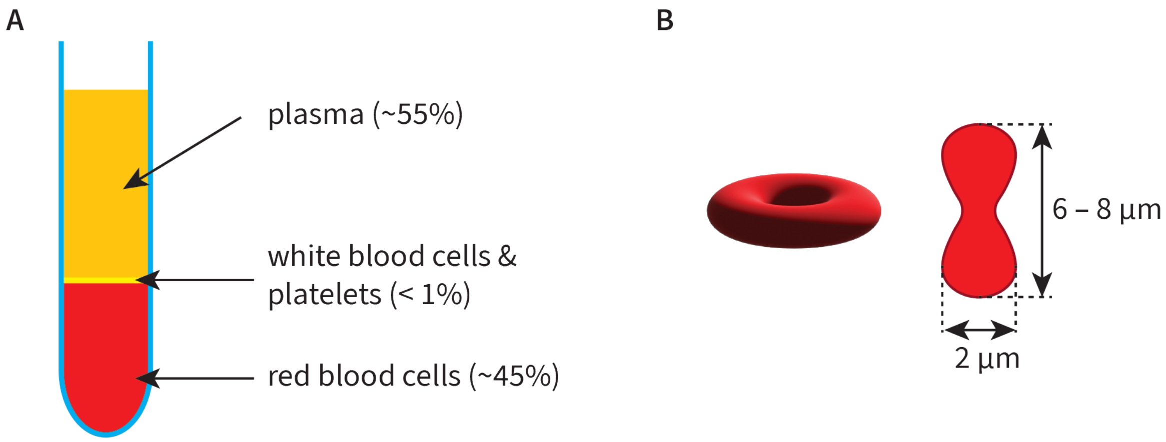


Figure 12.16. Blood composition. A, composition of blood by volume. Test tube shows approximate volume fractions of blood components following separation through centrifugation. B, shape of erythrocytes (red blood cells). Left – 3D view; right – cross section. Erythrocytes have the form of biconcave disks (thin center and thick margins). This provides a large surface for gas exchange and also allows the cell to deform if needed during circulation.

12 As we saw, distensibility is the ability of an object to get stretched, while elasticity is the ability of an object to recoil, returning to the original position after being stretched.

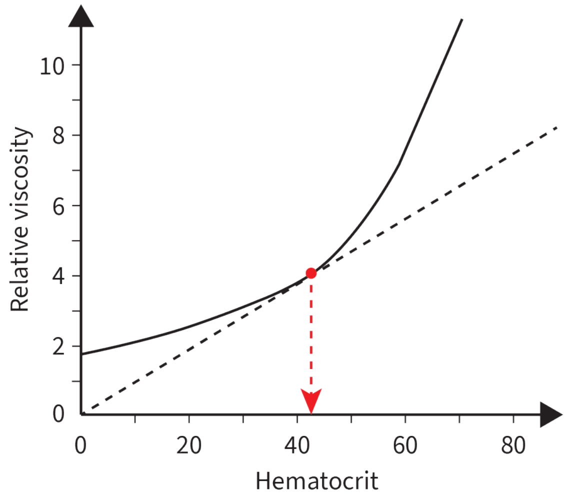


Figure 12.17. Relative blood viscosity increases almost exponentially with increasing hematocrit. The red arrow indicates the approximate optimal value of the hematocrit.

to flow decreases and so does the viscosity of blood. As a general rule, shear rate increases with increasing flow velocities and increases when vessel diameter decreases. Viscosity is highest in the veins, and lowest in the capillaries (see the Fåhræus–Lindqvist effect below).

Remember that, for the same volumetric flow, an increase in the resistance to flow leads to an increase in blood pressure, according to equation (12.25). In addition to that, resistance to flow is directly proportional to viscosity, as shown by equation (12.26). In conclusion, if, for whatever reason, the viscosity of arterial blood increases, the resistance to flow will increase and so will the arterial blood pressure.

Temperature is another factor that can alter the viscosity of blood. Normally, the temperature of blood should not change much in the body. However, if the extremities of the body are exposed to severe cold, the viscosity of blood increases, which means an increase in the resistance to flow and a decrease in volumetric flow. This can lead to severe hypoxia followed by necrosis of peripheral tissues (*frostbite*).

2.7. Flow velocity of blood

Due to the high viscosity of blood and its pulsatile ejection from the ventricles, but also due to the variable shape and branching of the blood vessels, the flow of blood is not steady but *pulsatile*. Under physiological conditions, at rest, the blood flow is turbulent only in the ascending part the aorta and pulmonary artery (where $Re > 3000$). In large arteries there is an intermediary kind of flow ($2000 < Re < 3000$) and in the rest of the blood vessels the flow is close to laminar. When physical effort is being done, blood flow can become turbulent in the entire aorta, in large arteries and the venae cavae. The distinction between laminar and turbulent flow has a high clinical relevance; **laminar flow is silent, whereas turbulent flow produces noises called murmurs**. These properties are used in the clinic to assess the arterial blood pressure using the auscultatory method (e.g. Korotkoff sounds are produced by the turbulent flow of blood through the partially constricted artery) or to diagnose pathological conditions like vessel stenosis, shunts, cardiac valvular lesions, etc.

The velocity of blood depends primarily on the level of branching. As blood passes from the large arteries (average velocity of 40 cm/s) into smaller arteries and arterioles, its velocity decreases. By the time it reaches the capillaries, the average velocity drops to ~1 mm/s. When blood moves from the capillaries into venules and, then veins, flow velocity then increases, though to overall lower values than in the arteries.

Remember the continuity equation (12.2) discussed in the beginning of this chapter? It states that the volumetric flow (Q) is constant along a tube and equal to the product of cross-sectional area (S) and the velocity of flow (v). When applying this equation to the circulatory system, we have to consider the overall cross-sectional area which is the sum of the individual cross-sectional areas of all the parallel blood vessels at a certain point of the circulatory system (Table 12.1). Therefore, although their radii are the smallest, capillaries have the highest level of branching and their overall cross-sectional area is about ~1000 times higher than that of the aorta.

2.8. The Fåhræus–Lindqvist effect

Apart from the hematocrit, the dimension of the blood vessel and the velocity of flow can also influence the viscosity of blood. The radius of a blood vessel can be an important determinant of blood viscosity. For radii larger than 1 mm, the

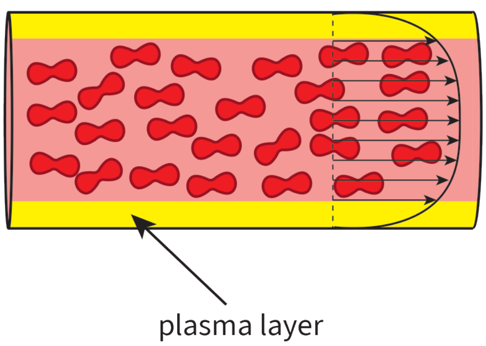


Figure 12.18. The Fåhræus–Lindqvist effect. In blood vessels of diameters lower than ~0.3 mm, red blood cells accumulate towards the center of the vessel, leaving only a layer of plasma at the walls. This lowers viscosity and increases velocity, flattening the flow profile.

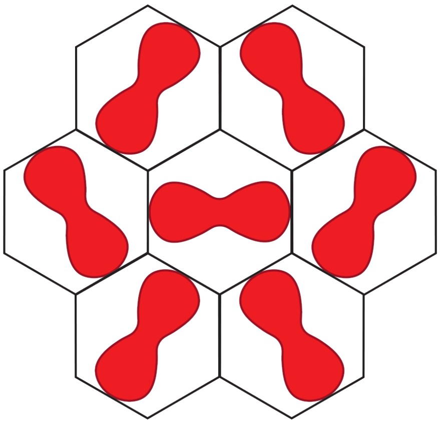


Figure 12.19. Ideal geometrical arrangement of red blood cells, corresponding to the highest concentration without deformation.

viscosity of blood is independent on the blood vessel radius. However, it decreases steeply at smaller radii. The reason for that is the tendency of red blood cells to move toward the center of the blood vessel (*axial accumulation*) when caught between two adjacent layers of plasma flowing at different velocities.

In blood vessels with diameters smaller than ~0.3 mm (arterioles, venules and capillaries), the *Fåhræus–Lindqvist effect* appears. The decrease in viscosity of blood in these vessels appears because the cells move away from the region near the wall (where the friction forces would be strongest) and their concentration is greater in the center. Red blood cells thus orient themselves with their long axis parallel to the axis of the blood vessel, in order to lower friction. A lower viscosity thus leads to a higher flow velocity. If you compare the flow profile of blood in capillaries (Figure 12.18) with the parabolic profile described by Poiseuille's law shown in Figure 12.5B, you should notice that the flow profile in the capillary is flatter towards the middle.

In even smaller capillaries, which have a diameter about the size of a red blood cell, we can no longer speak of friction between different layers of blood cells. Red blood cells flow one by one by rolling around the cytoplasm and even deforming if the diameter of the vessel is smaller than the diameter of the red blood cell. A layer of plasma separates the membrane of the red blood cell from the capillary wall.

2.9. Biophysical aspects of the pathology of blood circulation

▸ **Abnormal changes in hematocrit.** If we considered the red blood cells as being enclosed in hexagonal boxes (Figure 12.19), the hematocrit would be 58%, the maximal value at which the red blood cells do not deform. At higher values, the red blood cells change their shape due to elastic deformations and the viscosity of blood exceeds 0.06 P. In *polycythemia vera*, a type of blood cancer, the hematocrit can reach values of 70 – 80% due to excessive production of red blood cells. As already discussed, the viscosity of blood increases exponentially with increasing hematocrit (Figure 12.17). Higher viscosity means higher resistance to flow and a lower volumetric flow; obstructions of small blood vessels can appear which lead to local blockages of blood circulation.

▸ **Rigid narrowing of the vascular wall in atherosclerosis**. Atherosclerosis is a pathological condition in which LDL cholesterol[13] and triglycerides adhere to the arterial wall, macrophages are recruited and atheromatous plaques[14] form (Figure 12.20). Atheroma affects the mechanisms of local vasodilation and vasoconstriction of the vascular wall. The arterial lumen narrows, the blood flow velocity increases

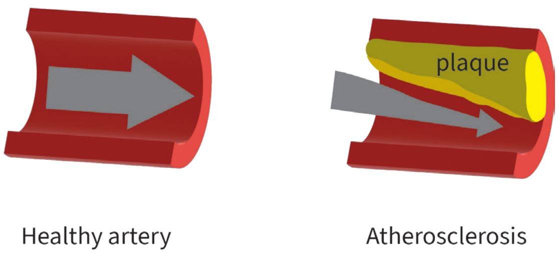


Figure 12.20. Narrowing of arteries in atherosclerosis. Arrows indicate blood flow.

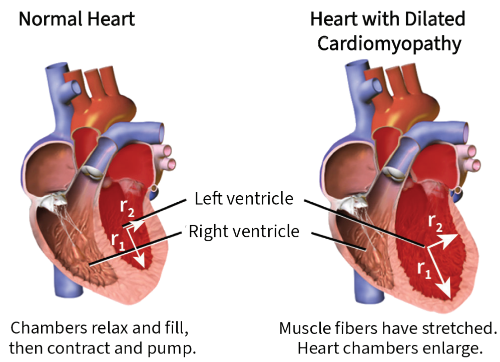


Figure 12.21. Dilated cardiomyopathy[15] resulting in enlargement of the left ventricle. White arrows indicate the main radii of curvature of the left ventricle, r_1 and r_2.

13 LDL = low density lipoprotein.

14 An atheroma, or atheromatous plaque, is an abnormal accumulation of material in the arterial lumen.

15 Modified from copyright-free image provided by Blausen.com staff (2014). "Medical gallery of Blausen Medical 2014". WikiJournal of Medicine 1 (2). DOI:*10.15347/wjm/2014.010*. ISSN 2002-4436.

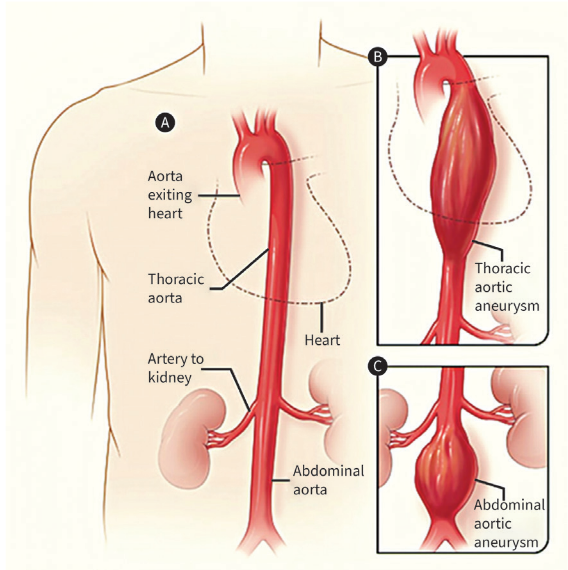


Figure 12.22. Aortic aneurysms.[16] A – normal aorta, B – thoracic aortic aneurysm, C – abdominal aortic aneurysm.

and the local pressure will drop. This decrease in blood pressure might affect the perfusion of organs irrigated by that artery. The high velocity and the less smooth lumen surface (at the stenotic point) favor the appearance of turbulences which cause murmurs.

In later stages of the condition, the artery might become completely blocked, leading to ischemia. If the blocked artery is a coronary artery, this will cause myocardial infarction (heart attack). If a blood vessel in the brain is blocked, this will cause a stroke.

- **Changes in the size of the heart**. In *dilated cardiomyopathy*, the chambers of the heart are enlarged and the main radii of curvature are also larger (Figure 12.21). The shape of the left ventricle becomes more spherical and its walls become thinner. Again, Laplace's law, as shown in equation (12.19), is useful in understanding the consequences. The myocardium has to make a greater effort in order to produce, through contraction, a stronger tension in the cardiac walls for ensuring a normal systolic pressure. Eventually, dilated cardiomyopathy progresses to heart failure.
- **Aneurysm.** Large arteries need to withstand very large wall tensions which are proportional to the blood pressure and the arterial radius. This is shown, as well, by the law of Laplace, according to equation (12.20). Remember that elastic fibers (passively) and smooth muscle fibers (actively) contribute to the vascular wall tension. Elastic fibers are important not only in giving the vascular wall resistance to high transmural pressure, but also for ensuring a graded response when smooth muscle tone changes.

An imbalance between the passive and the active contributors to the vascular wall tension can lead to pathological situations. For example, if the elastic content is reduced or damaged, vessels tend to become larger (dilate), under the action of blood pressure (Figure 12.22). Bulging of the artery leads to an increase in the wall tension. In this way, the process is progressively amplified and an **aneurysm** develops. If left untreated, the wall of the blood vessel becomes progressively weaker and the aneurysm eventually ruptures.

16 Modified from a public domain image provided by the National Heart, Lung, and Blood Institute; National Institutes of Health; U.S. Department of Health and Human Services.

REFERENCES

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Boron, W. F., & Boulpaep, E. L. (2017). *Medical Physiology* (3 ed.). Philadelphia: Elsevier.

California Institute of Technology, Gottlieb, M. A., & Pfeiffer, R. (2013). *The Feynman Lectures on Physics*. Retrieved from https://www.feynmanlectures.caltech.edu/

Franklin, K., Muir, P., Scott, T., & Yates, P. (2019). *Introduction to Biological Physics for the Health and Life Sciences*: Wiley.

Gordon Betts, J., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., . . . DeSaix, P. (2022). *Anatomy and Physiology 2e*. Retrieved from https://openstax.org/books/anatomy-and-physiology-2e/pages/1-introduction

Guyton, A. C., & Hall, J. E. (2005). *Textbook of Medical Physiology. Eleventh Edition*. Philadelphia: Elsevier.

Williams, B., Mancia, G., Spiering, W., Agabiti Rosei, E., Azizi, M., Burnier, M., . . . Group, E. S. D. (2018). 2018 ESC/ESH Guidelines for the management of arterial hypertension: The Task Force for the management of arterial hypertension of the European Society of Cardiology (ESC) and the European Society of Hypertension (ESH). *European Heart Journal, 39*(33), 3021-3104. doi:10.1093/eurheartj/ehy339

Yartsev, A. Deranged Physiology. Retrieved from https://derangedphysiology.com/main/home

CHAPTER 13

MEDICAL IMAGING TECHNIQUES

Prerequisite knowledge

- Ionizing radiation. Dosimetry
- Interaction of ionizing radiation with living organisms
- Radioactive isotopes
- Types of radiation. EM and acoustic waves

1. GENERAL CONCEPTS OF MEDICAL IMAGING

Medical imaging allows the visualization of the structure and assessment of the function of organs and tissues inside the body, for the purpose of diagnosis and treatment. This is accomplished by either directing some type of radiation (ionizing or non-ionizing) towards the body and following the interaction of this radiation with the tissues and organs or by detecting the radiation naturally produced by our bodies. The final product is a *medical picture* showing the region of interest.

A huge advantage of medical imaging is its ability to generate images of the interior of the body without requiring surgery. This makes obtaining the right diagnosis or following the evolution of a condition much faster, and causes less discomfort to the patient. However, the drawback is that there is a risk associated with some medical imaging techniques, especially if they employ ionizing radiation. We will, therefore, emphasize the potential risk of harm that each technique represents to the patient when discussing it. To make this process easier, we will refer to this risk as the **invasiveness** of the technique.

In a strict definition that you might find in a dictionary, invasive procedures are those that break the skin or insert a foreign object into the body. While this takes into account the degree of discomfort caused to the patient, it says nothing about the risk of harm of the respective procedure. We will, therefore, extend the definition of invasiveness and say as follows:

An invasive technique is one that inserts any kind of material (substances, needles, etc.) or any kind of energy (X-rays, acoustic waves, γ radiation, etc.) into the body.

We include into our extended definition of invasiveness the notion of potential of harm. A technique will be less invasive if it is less potentially harmful to the patient. We can, thus, classify the techniques that will be presented in this chapter as:

- **non-invasive**, that have zero potential of causing harm to the patient: **thermography**;
- **minimally invasive**, that have a low potential of causing harm to the patient: **MRI, ultrasonography**;
- **highly invasive**, that carry a high potential risk to the patient: **all techniques involving ionizing radiation – radiography, CT, scintigraphy, PET, SPECT.**

2. RADIOGRAPHY

2.1. Characteristics of X-rays

As we have seen in the chapter on Photobiology, we call ionizing radiation any radiation that is capable of ionizing atoms or molecules by extracting electrons from their structure (Figure 8.5). **Radiation with energies above 10 eV is considered ionizing radiation**.

Medical imaging was essentially born with the discovery of X-rays (EM radiation with wavelengths of 10 pm – 10 nm) by Wilhelm Röntgen in 1895, for which he was awarded the first Nobel Prize in Physics in 1901. Since that time, other imaging techniques have emerged, either using X-rays or other types of radiation. The discovery of X-rays was a tremendous leap for both physics and medicine. In Röntgen's honor, X-rays are also called Röntgen radiation[1]. One of Röntgen's early observations was that the newly-discovered type of radiation (which he named X, denoting them as "unknown") was able to penetrate the human body and project an image of the skeleton. One of the famous first radiographic pictures he took (Figure 13.1) was of his wife's hand.

2.2. Production of X-rays

Radiography uses X-rays with a continuous

1 In German-speaking countries the name for X-rays is *Röntgenstrahlung*, while radiography is simply called *Röntgen*.



Figure 13.1. One of the first radiographic images[2]. Pictured is the hand of Anna Bertha Ludwig, Wilhelm Röntgen's wife. A ring she carried on her finger can also be easily seen. In this radiographic image, bones appear black because the early X-ray images were reproduced after developing the photographic film in the positive mode.

spectrum produced as *braking radiation*[3] (Figure 13.2). This is produced following the braking of accelerated electrons in a device called an *X-ray tube*. X-ray tubes used in the present are all based on an initial design by William Coolidge, who invented the *Coolidge tube* in 1913.

The **X-ray tube** is a glass tube with the interior kept at a high vacuum. Two electrodes (metal pieces) are present in the tube, an *anode* and a *cathode*. The cathode is a tungsten filament that is heated by passing an electric current through it. This will cause the cathode to emit electrons through *thermionic emission*. These electrons are accelerated and directed towards the anode by applying a high potential difference between the cathode and the anode (on the order of tens of kV).

The anode is made out of a heavy metal (tungsten or molybdenum). When the accelerated electron beam reaches the anode, the production of X-rays will occur: an electron that arrives in close proximity to a nucleus of the metal making up the anode will lose most of its energy, which is emitted as an X-ray photon (the braking radiation or *Bremsstrahlung*). Thereby, the electron is slowed down.

Besides braking radiation, X-rays can also be emitted as characteristic X-rays. These are produced when a nucleus is bombarded with high energy particles, causing electrons in their inner electron shells to be ejected, leaving vacancies. These vacancies are filled by other electrons at higher energy levels "dropping down" to the unfilled vacancy, thereby releasing X-ray photons. Characteristic X-rays have particular energies depending on the energy levels in the target atom (the atom bombarded with electrons). These are not relevant in radiography, but are used in science, and biophysics in particular. For example, characteristic X-rays can be directed at protein crystals in order to determine their 3D structure through X-ray diffraction.

The process of bombarding the anode with electrons is highly damaging to the anode, which is quickly heated up and deteriorated. To slow down the deterioration of the anode, it can be rotated (Figure 13.2) to expose different areas to the electron beam, and it can be cooled down. However, the anode is still damaged over time and the X-ray tube has to be replaced periodically.

Depending on area being imaged and the desired image contrast, the X-ray beam produced by the X-ray tube can be adjusted according to two

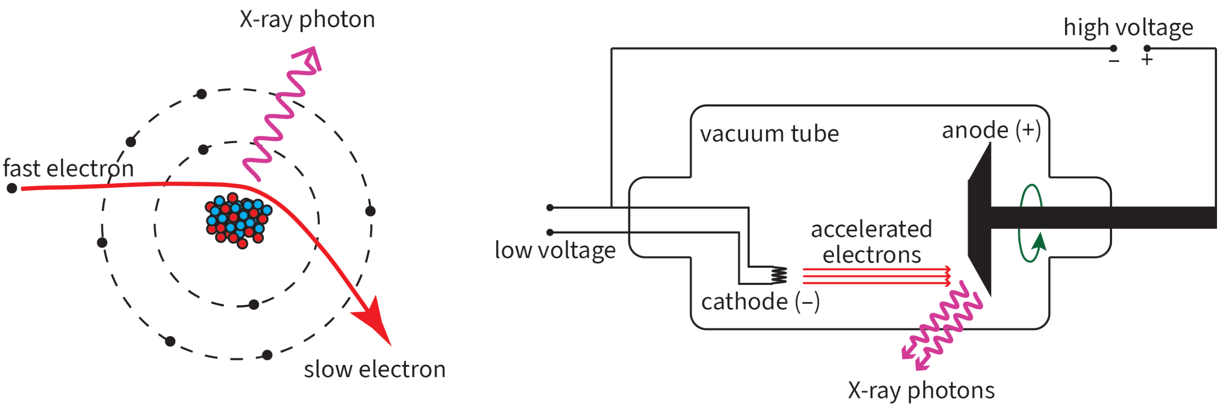


Figure 13.2. Production of X-rays in radiography. Left panel: the production of braking radiation is achieved when a fast electron is slowed down in the proximity of a heavy nucleus. Right panel: an X-ray tube (rotating anode tube) used for the production of X-rays. Low voltage refers to the circuit heating up the cathode. High voltage refers to the circuit accelerating the produced electrons. Detailed description is in the text.

2 Public domain image available on Wikimedia Commons (https://commons.wikimedia.org/wiki/File:First_medical_X-ray_by_Wilhelm_R%C3%B6ntgen_of_his_wife_Anna_Bertha_Ludwig%27s_hand_-_18951222.gif).

3 The German name for braking radiation is also commonly used: *Bremsstrahlung*.

parameters:

- The beam current, which shows how many electrons flow from the cathode to the anode each second. If multiplied by the time of exposure, this will be a measure of the **quantity** of X-rays that the patient receives. Radiologists refer to this as the milliampere-seconds “**mAs**” or just “**mA**”. A higher mAs means that the patient will receive a higher dose;
- The voltage at which the electrons are accelerated, which controls the energy of the produced X-ray photons. Radiologists refer to this as the kilovoltage “**kV**” or the peak kilovoltage “**kVp**”. Higher kV is used for imaging thicker anatomical regions, as X-rays produced at higher kV are more penetrating. However, increase of kV leads to increase of Compton scattering, and thus decrease of image contrast. Thus, kV shows the **quality** of the X-rays produced. Increasing kV also increases the dose.

2.3. Interaction of X-rays with tissue

Why could Röntgen detect images when directing X-rays at tissue (Figure 13.1)? We have already established in the Radiobiology lecture that X-rays and γ rays are highly penetrating. As an X-ray beam passes through tissue, part of its energy will be absorbed by the tissue – this is called *attenuation*. Attenuation results from either scattering or absorption of the X-rays, according to the mechanisms described below.

The interaction of X-rays with matter is classified into 4 categories, 2 types of scattering and 2 types of absorption that we will briefly describe (Figure 13.3):

- *elastic (Rayleigh) scattering*: low energy X-ray photons are deviated by an atom without a change

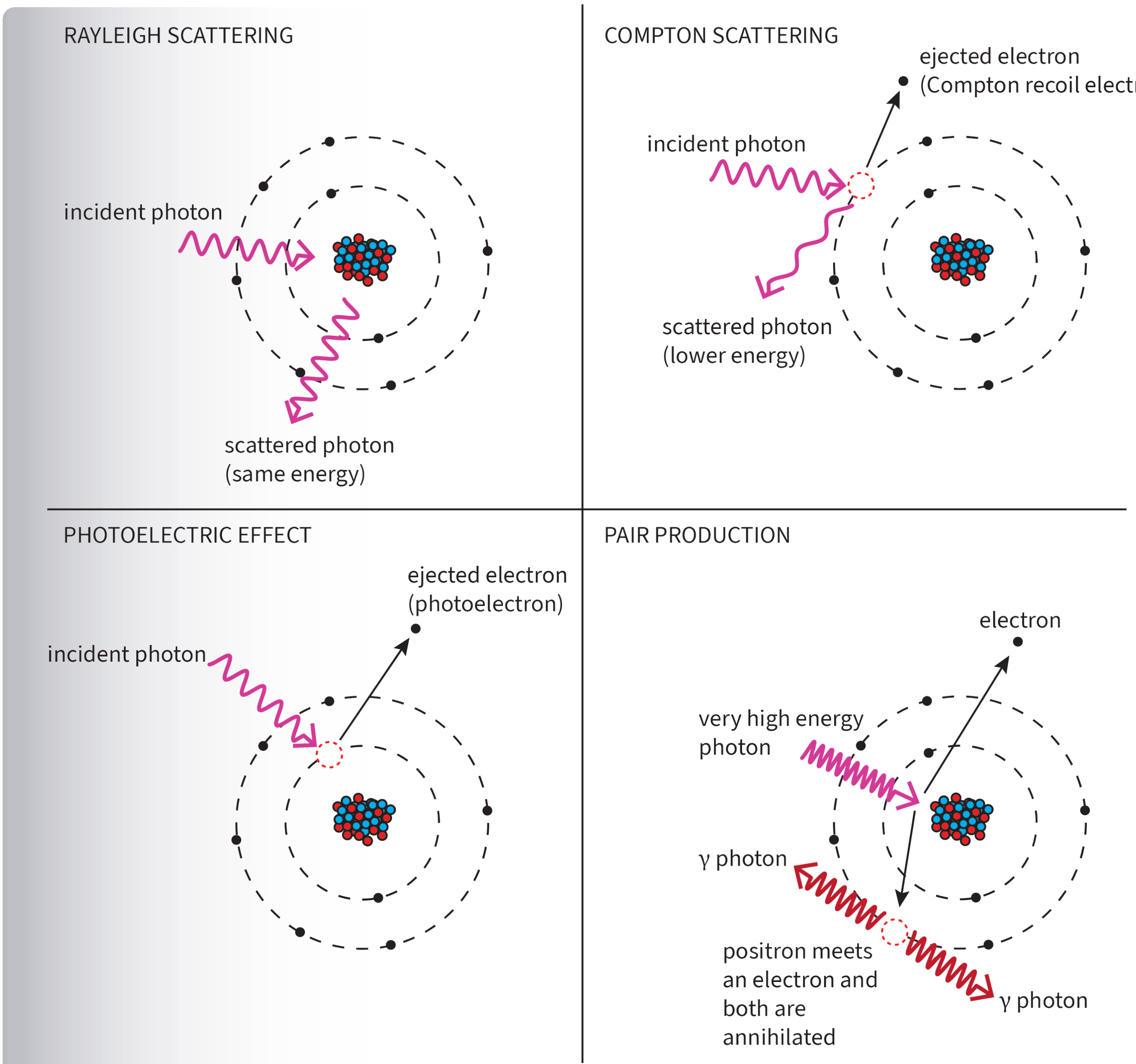


Figure 13.3. Interaction of X-rays with matter. A detailed description of each process is provided in the text. Dashed circles represent the removal or disappearance of an electron. At the energies used for the X-ray beam in radiography, Compton scattering and the photoelectric effect are the predominant processes in tissue.

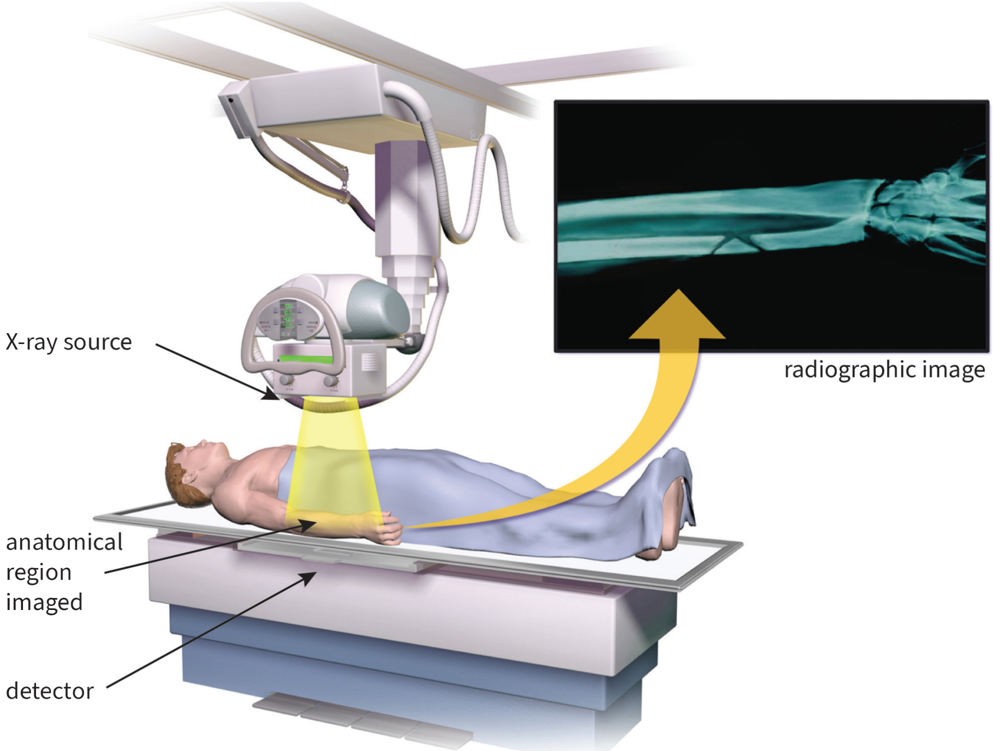


Figure 13.4. Radiography of a patient's forearm[4].

in the energy of the atom or of the photon. This affects image quality, but has low probability at the energies used in radiography;

- *Compton scattering*: an X-ray photon imparts a portion of its energy to a valence (outer) electron in an atom, and the electron is ejected from the atom (the atom is ionized). The X-ray photon is deviated and will have a higher wavelength, as it transferred part of its energy to the electron. The ejected electron (called *recoil electron*) will excite or ionize surrounding tissue, contributing to the patient's dose. The probability of Compton scattering increases with the density of the material. Scattered X-rays lower image quality and contrast. This is the major interaction of X-rays with soft tissue or air;
- *the photoelectric effect*: an X-ray photon transfers all its energy to an electron, thereby ionizing the atom. The ejected electron will carry all of the energy of the absorbed X-ray (no other X-ray is emitted, as in Compton scattering). The probability of this effect increases with the atomic number of an element. This is the major interaction of X-rays with contrast media or bones;
- *pair production*: at very high energies (>1 MeV), an X-ray photon interacts with the nucleus of the atom and its energy is transformed into an electron-positron pair. These particles will contribute to ionization and excitation of the material. The positron will travel in tissue until it meets another electron, at which point an annihilation event will occur (see the Radiobiology chapter), whereby the pair will disappear, releasing two γ photons. This phenomenon is not encountered at energies used in diagnostic X-rays.

2.4. Basis of radiographic imaging

Radiography is the imaging technique that produces a flat (2D) picture showing the differential absorption of X-rays while passing through a certain anatomical region. In order to obtain a radiographic image, X-rays are directed at a particular area of the patient's body. These will be absorbed differently by tissue depending on the characteristics of the tissue. The portion of X-rays that passes through the patient will be captured on a detector placed behind the patient's body. The detector can be either a photographic film, or, in modern devices, a digital detector. The negative image of the film is used in radiography. **Thus, portions of the film exposed to a higher amount of X-rays** (coming from tissue that attenuated very little) **appear as black, whereas portions of the film reached by the lowest amount of X-rays** (coming from highly attenuating tissue) **will be white.**

We obtain in this manner 2D images that represent the projection of the anatomical regions through which the X-ray traveled (Figure 13.4). Consequently, the technique is also

4 Modified from copyright-free image by Bruce Blaus, available at *https://commons.wikimedia.org/wiki/File:X-Ray.png*

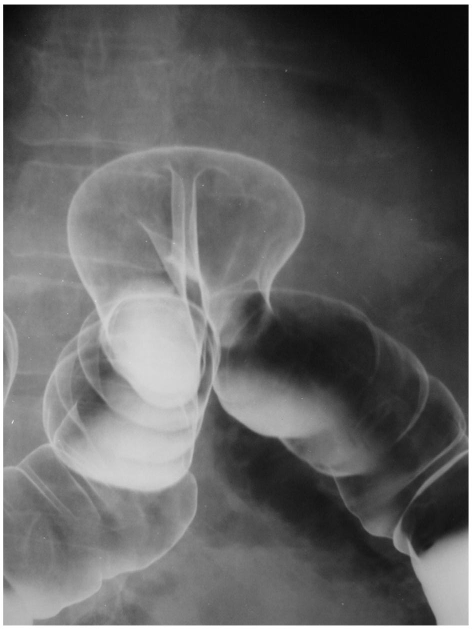
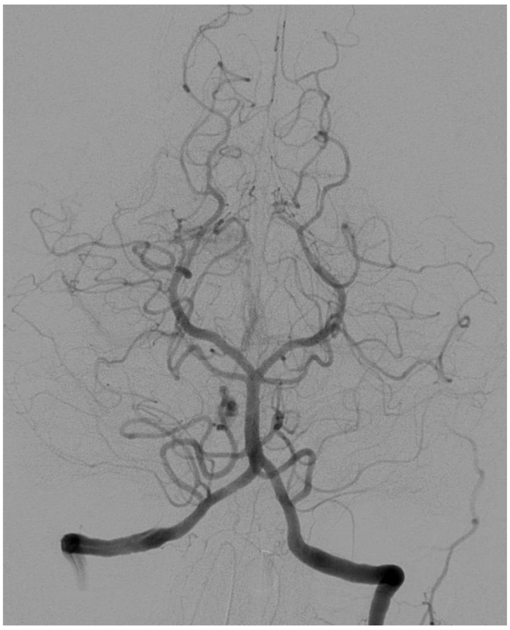



Figure 13.5. Contrast agents. Left[5]: double contrast radiography of a colonic herniation using barium sulfate (radiopaque – white) and air (radiolucent – black) as contrast agents. Right[6]: cerebral angiography using an iodine-based compound (radiopaque). As the right image is fluoroscopic, the image color is inverted and X-ray attenuating structures are shown as black (see below).

called *projectional radiography* or *conventional radiography*.

The radiographic image is produced as a consequence of the differential absorption of X-rays. Depending on their composition, some tissues will absorb more and some will absorb less. We can say that each tissue has a particular *radiological density*, which mainly depends on the density of the tissue and the atomic number of the atoms that make up the tissue.

There are four basic radiological densities:

- bone: it attenuates the highest amount of X-rays through the photoelectric effect and will appear as white (*radiodense* or *radiopaque*);
- soft tissue: fluids and soft tissues have approximately the same densities. These will appear as lighter shades of grey;
- fat: it is less dense than organs and will appear as darker shades of grey;
- air: it attenuates very little, and it appears as black (*radiolucent*).

An additional, fifth type of density may appear in the bodies of some patients: metal, which is more radiodense than bone. This may come, for instance, from orthopedic implants.

2.5. Common problems in X-ray imaging. Radiocontrast agents

We will limit ourselves to listing in brief some common issues that affect the radiographic image.

One of the most common problems encountered in radiography is the **superposition** of anatomical structures inside the body. When the X-ray beam travels through several structures before reaching the detector, it will be attenuated by all of these. This can create an appearance of an increased density and also eliminates information regarding depth. Superposition can make some anatomical structures impossible to distinguish from one another. Excellent anatomical knowledge is required when analyzing radiographic images.

The **size** of an anatomical structure on the X-ray image depends on the proximity to the X-ray source. Thus, structures which are closer to the source will appear as larger on the radiograph. This can easily be seen when comparing the size of the heart on an AP (antero-posterior, the source is anterior and the detector is posterior) projection compared to a PA (postero-anterior, the source is posterior, and the detector is anterior) projection: as the heart is an anterior structure in the chest, it will appear larger in the AP projection than in the PA projection. We see, thus, that position of the patient relative to the X-ray source and detector matter a lot in radiography! Therefore, *standard positions* are used when imaging different body parts.

Scatter of X-rays will negatively influence the quality of the X-ray image. X-rays that are scattered by tissue change direction in the body, but

5 Public domain image by Nevit Dilmen, retrieved from *https://commons.wikimedia.org/wiki/File:Colonic_Herniation_08787.jpg*

6 Public domain image by Lipothymia, retrieved from *https://commons.wikimedia.org/wiki/File:Cerebral_angiography,_arteria_vertebralis_sinister_injection.JPG*

might still arrive at the X-ray detector, covering both the dark and white parts of the image. This will cause the image to be noisy.

We call **contrast** the difference in brightness between two adjacent areas of the radiograph. Contrast can be increased by lowering kV. *Contrast agents* (Figure 13.5) can also be employed in order to increase contrast. These are chemical compounds containing elements of a higher atomic number that attenuate radiation more than soft tissue. Contrast agents allow the increase of contrast for regions of the body where they are distributed. Ideally, a contrast agent should be non-toxic, physiologically inert, and should be quickly and fully be eliminated from the body. An example of a commonly used contrast agent is barium sulfate (Figure 13.5, left panel), which can be swallowed by the patient in order to allow enhanced visualization of the patient's gastrointestinal tract. For imaging soft tissues and the circulatory system, iodine compounds are used (Figure 13.5, right panel).

2.6. Fluoroscopy

Radiography results in single images, essentially snapshots taken at a particular point in time. If real time X-ray images need to be obtained (for example, for monitoring complex interventions), the technique is called *fluoroscopy*.

Classically, fluoroscopy was monitored on a fluorescent screen (hence the name), which emits visible light after absorbing X-rays. Therefore, a fluoroscopic image is a positive image, unlike the negative image of radiography: bones appear as black and air appears as white (Figure 13.5, right panel). Note that, currently, digital detectors are used and the image can be shown in any color convention desired (either radiographic or fluoroscopic). However, the convention in fluoroscopy has been kept for historical reasons.

Fluoroscopy exposes the patient to a much higher dose than classical radiography, due to the necessity of keeping the X-ray beam turned on for as long as the intervention requires it. Unlike classical radiography, where in most situations medical staff is outside the room while the radiograph is taken, fluoroscopy can expose medical personnel to X-rays scattered from the patient's body. Lead aprons are worn by all persons in the room, except the patient.

Examples of fluoroscopic procedures are: barium swallow and enemas for the gastrointestinal tract, angiography (imaging of the lumen of blood vessels), guided stent placement, etc.

2.7. Doses in radiography and fluoroscopy

The effective dose delivered to the patient in radiography depends on the tissue imaged as well as the characteristics of the X-ray beam (mAs and kV). The dose in classical radiography can range from a few μSv up to tens of μSv, while in fluoroscopy the doses can be much higher (a few mSv). For examples of doses, refer to Table 9.6 in the chapter on Radiobiology.

3. COMPUTED TOMOGRAPHY (CT)

3.1. Basis of CT

We mentioned that one of the major issues physicians have to contend with in radiography is superposition. Thus, attenuation of the X-ray beam is cumulative and performed by all anatomical structures in the path of the beam. This makes interpreting some radiographic images challenging, as well as making it impossible to obtain good images from certain orientations.

To overcome this limitation, an engineer called Godfrey Hounsfield came up with the idea of taking multiple radiographic images of an object from different angles in order to create a more detailed picture of the object that could be viewed in slices. Using the mathematical calculations published earlier by physicist Allan McCormack, Hounsfield built the first CT machine in 1971. In 1979, the two were jointly awarded the Nobel prize in Physiology or Medicine for this achievement.

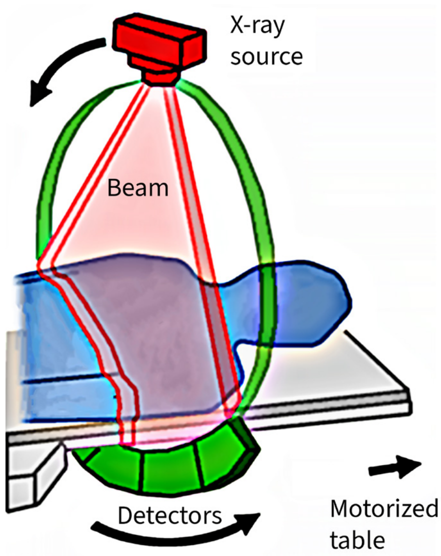


Figure 13.6. Schematic of a modern CT machine[7]. Description is in the text.

7 Modified from a public domain image provided by the US Food an Drug Administration (https://www.fda.gov/radiation-emitting-products/medical-x-ray-imaging/what-computed-tomography).

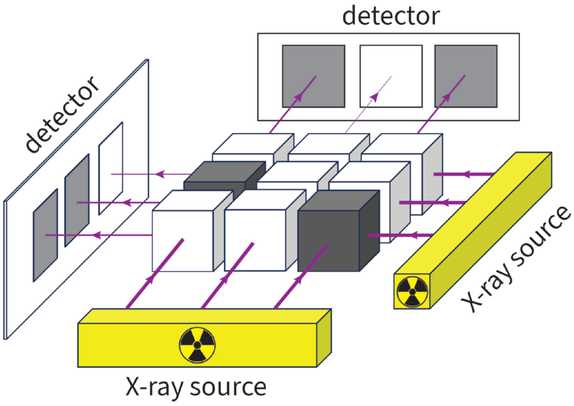


Figure 13.7. Schematic mechanism of a CT algorithm. A volume of tissue containing 9 voxels in a 3x3 grid is imaged. 7 of the voxels are radiopaque (white), 2 are radiolucent (grey). In such a simple arrangement, exposing the volume to X-rays from 2 orthogonal directions allows determination of the position of the radiolucent voxels. For a volume containing more voxels, solving the density of each voxel will require more images (more directions of exposure). The thickness of the lines suggests the differential attenuation.

We call *computed tomography* (CT) the radiographic technique in which X-rays are directed at the patient in a narrow beam that is quickly rotated around the body. This results in the production of hundreds or thousands of radiographic images that are then processed by a computer in order to generate cross-sectional images called *slices*.

A schematic of a modern CT machine is shown in Figure 13.6. The X-ray source and the detectors are placed on a circular structure that can rotate around the patient. A fan-shaped X-ray beam is directed at the patient and a different X-ray image is detected by each detector. Turning the system of X-ray source and detectors allows imaging the respective slice of the patient's body from all possible angles. In order to image a different slice, the table on which the patient sits can be moved.

Table 13.1. Hounsfield unit (HU) values for different tissue types[8]. Right column shows the correspondence between the HU and the displayed shade of gray for the respective pixel in the CT image.

Tissue	HU	Grayscale
Dense bone	+1000	
Contrast agent	+130	
Muscle / soft tissue	+40	
Water	0	
Fat	–100	
Air	–1000	

8 According to Kissane, J., Neutze, J. A., & Singh, H. (2020). *Introduction to Radiology Concepts*. In J. Kissane, J. A. Neutze, & H. Singh (Eds.), *Radiology Fundamentals: Introduction to Imaging & Technology* (pp. 11-14). Cham: Springer International Publishing.

In CT, the volume of tissue that needs to be imaged is divided into *voxels* (volumetric pixels): tiny elements of volume (3D). These are analogous to the pixels for a 2D surface with an added dimension of thickness. Based on the large number of radiographic images produced, each voxel is assigned a certain radiodensity. A schematic for how that process would work for a small number of voxels is shown in Figure 13.7.

Radiodensity is expressed through the form of a quantitative scale named the Hounsfield scale. In the Hounsfield scale, each voxel is assigned a *Hounsfield unit* (HU, called also *CT number*), that corresponds to its attenuation of X-rays (Table 13.1). In the final CT image, the HU is represented as the color of a pixel on a grayscale.

The attenuation coefficient of water is assigned a 0 HU value while air is assigned –1000 HU. The HU value for a given voxel is defined as:

$$HU = 1000 \cdot \frac{\mu - \mu_{water}}{\mu_{water} - \mu_{air}} \qquad (13.1)$$

where μ is the attenuation coefficient of the voxel, while μ_{water} and μ_{air} are the attenuation coefficients for the respective substance.

3.2. Advantages of CT. Examples of CT images

CT allows us to **image axial slices** of the patient's body (imagine slicing the patient's body in the axial plane like you would be slicing bread) that cannot be obtained using classical radiography. CT reconstruction almost completely **eliminates the problem of superposition** that is found in classical radiography. Furthermore, the **contrast is much improved in CT**, allowing the distinction between tissues that have close HU values. Another big advantage of CT is the **fast acquisition time**. This allows CT to be used also in emergency situations.

Figure 13.8 shows an example of a CT investigation performed on a COVID-19 patient.

3.3. Disadvantage of CT. Doses in CT

The major disadvantage of CT is the **dose that the patient receives**. Typically, the effective doses are on the order of several mSv and can be more than what the patient is exposed to in an entire year from the natural background radiation. For example, the effective dose for the patient in Figure 13.8 was of ~2.6 mSv for a chest CT scan using 30 mAs and 120 kV. By comparison, the single radiographic exposure in the left panel imparted an effective dose of 0.06 mSv. We have provided more examples of doses in CT in the chapter on Radiobiology (Table 9.6). You can see there that some CT procedures can expose the patient to

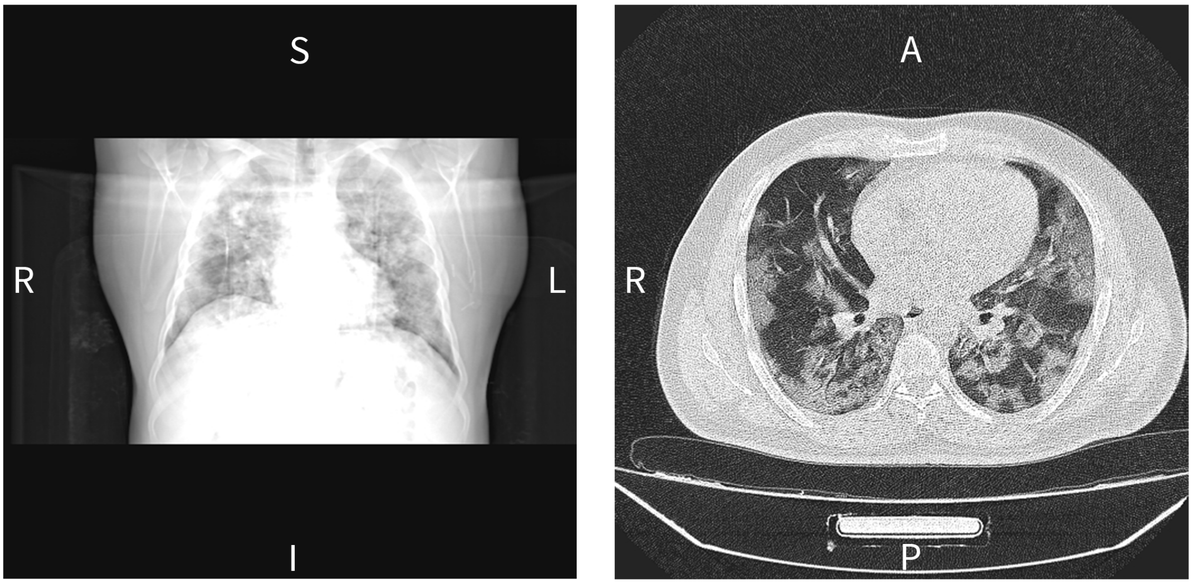


Figure 13.8. X-ray imaging of a COVID-19 patient[9]. The patient was a 46-year-old male. Left – frontal radiograph. Right – axial slice of the chest following CT. Opaque regions (called ground-glass opacities) characteristic of COVID-19 can be observed in the lungs. R – right, L – left, S – superior, I – inferior, A – anterior, P – posterior.

effective doses of more than 10 mSv.

We have seen that exposure to ionizing radiation even at low doses has an associated cancer risk. A rough estimate is that out of every 1000 patients exposed to a CT that imparts a dose of ~10 mSv (such as that for a full body scan), 1 patient will develop a cancer resulting from that investigation in the future. While these numbers are extrapolated, not directly measured, as we have discussed in the Radiobiology chapter, you as future doctors should be well aware that CT is risky for the patient and should only be performed if there is a clear necessity. If there is no hurry, alternative techniques can be used that do not expose the patient to ionizing radiation, such as MRI.

4. MAGNETIC RESONANCE IMAGING (MRI)

4.1. Nuclear spins. The Larmor frequency

Magnetic resonance imaging (MRI) is an imaging technique that provides high contrast images of tissues without exposing the patient to ionizing radiation. MRI was developed by Peter Mansfield and Paul Lauterbur in the 1970s, a discovery for which they were jointly awarded the Nobel Prize in Physiology or Medicine in 2003.

MRI is based on the principle of nuclear magnetic resonance (NMR). Note that detailed understanding of the principles of NMR requires good knowledge of quantum mechanics. As this is not part of your curriculum as future doctors, we will use a simplified explanation of the basic interactions that is mainly based on classical mechanics.

You might have learned in high school that electrons have an associated *spin* number. In classical mechanics you can imagine this number as describing how the electron rotates around its axis. While this simple image has been proven wrong, it is good enough for our text. For an electron, the spin number is either ½ or –½. Thus, when two electrons of opposite spin are paired, their spins "cancel each other out" and the total spin of the pair is 0 (Figure 13.9).

Moving electric charges produce magnetic fields. Thus, the way in which electrons are distributed will cause a chemical species to have certain magnetic properties. This is described by a quantity called *magnetic susceptibility*. Depending on how a chemical species (atom, ion or molecule) interacts with external magnetic fields, it

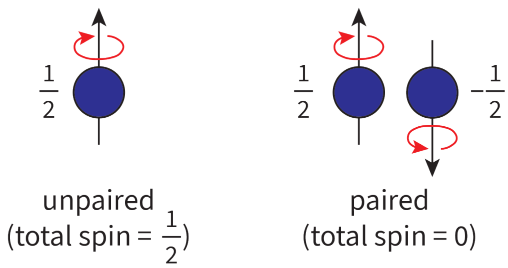


Figure 13.9. Total spin. An unpaired electron has a half-integer spin, while an electron pair has zero spin.

9 Public domain images provided by Mostafavi, S. M. (2021). *COVID19-CT-Dataset: An Open-Access Chest CT Image Repository of 1000+ Patients with Confirmed COVID-19 Diagnosis*. Retrieved from: *https://doi.org/10.7910/DVN/6ACUZJ*

Table 13.2. Nuclei and their total spins. The total abundance in nature of each isotope is given as a percentage.

Nucleus	Total spin	Isotope abundance (%)
$^{1}_{1}H$	½	99.99
$^{2}_{1}H$	1	0.01
$^{12}_{6}C$	0	98.93
$^{13}_{6}C$	½	1.07
$^{14}_{7}N$	1	99.6
$^{16}_{8}O$	0	99.76
$^{31}_{15}P$	½	100

can be classified as:

- *diamagnetic*: it is repelled by an external magnetic field. This occurs in chemical species with paired electrons (e.g., water);
- *paramagnetic*: it is attracted by an external magnetic field. This occurs in chemical species with unpaired electrons (e.g., oxygen);
- *ferromagnetic*: it is highly attracted by an external magnetic field (e.g. iron, cobalt).

We said that we will explain nuclear magnetic resonance, so why did we start out with electrons? We did it to simply serve as an analogy, especially since pairing of electrons is something you might have heard about in high school (and, if not, at least in our chapter on the properties of the water molecule).

Let's now look at the nucleus. The neutrons and protons that make up the nucleus of an atom also have an associated spin that can either be ½ or –½. As for the electrons, spin pairs can form which essentially "cancel out" their spins. Thus, the nucleus overall will have a total spin that depends on the number of protons and neutrons (on the Z and A numbers).

A nucleus that has a non-zero spin has a non-zero *magnetic moment*. In simple terms, we can say that such nuclei can act as tiny magnets and can be influenced by external magnetic fields. Nuclei that have a non-zero magnetic moment can be used in MRI.

Generally, we can say that:

- if the number of protons is even and the number of neutrons is even, the total spin is 0;
- if the number of protons is odd and the number of neutrons is even, or the other way around, the total spin is a half integer (½, 3⁄2, etc.);
- if the number of protons is odd and the number of neutrons is odd, the total spin is an integer number.

Examples of nuclei and their total spins are given in Table 13.2. Thus, we can conclude that we can visualize ^{1}H or ^{31}P using MRI, but not ^{16}O.

In the clinic, MRI is almost always performed to visualize the distribution of ^{1}H, as our body is mainly made out of water and hydrogen is

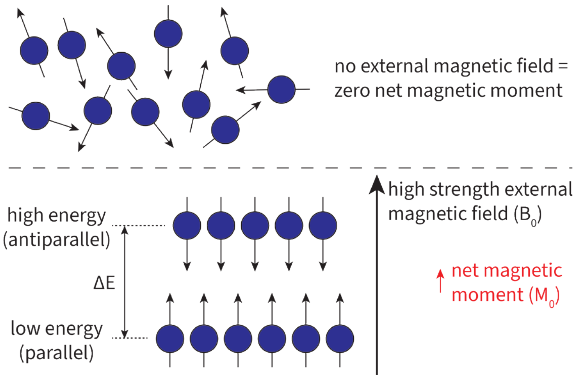


Figure 13.10. Behavior of protons in the absence or presence of an external magnetic field. Top, in the absence of an external field, the proton magnetic moments are oriented randomly, producing a total magnetic moment of zero. Bottom, when placed in an external magnetic field of strength B_0, protons orient their spins either parallel or antiparallel with the field. Slightly more protons are in the low energy (parallel) state, causing the appearance of a net magnetic moment (M_0) aligned parallel to the external field. ΔE represents the difference in energy between the parallel and the antiparallel states.

abundant in all of our tissues. Furthermore, ^{1}H makes up almost all (99.9%) of the total hydrogen in our body. Other nuclei that can be used in very specific applications are ^{13}C, ^{19}F, ^{23}Na or ^{31}P.

The ^{1}H nucleus is extremely simple, as it is made up of just one proton. Let's imagine a volume containing a certain number of protons (^{1}H nuclei). When not exposed to an external magnetic field, the magnetic moments of these nuclei are oriented randomly (Figure 13.10). If we add up all these tiny vectors, we will get a total magnetic moment of zero.

An MRI machine needs a very powerful magnet in order to function. These magnets have magnetic field strengths[10] of 0.3 to 4 T (by comparison, the magnetic field of the Earth has a strength of 25 – 65 µT). When placing protons in this strong magnetic field, a dynamic equilibrium state forms and their spins will align themselves with the direction of the field (Figure 13.10) in one of two possible orientations:

- a *low energy* state, where spins are parallel with the external field;
- a *high energy* state, where the spins are antiparallel with the external field.

Overall, just a few more protons will be in the low energy state, about 3 in 1 million for a field strength of 1 T. While this number might seem small, in a typical MRI voxel volume, this will correspond to ~$3 \cdot 10^{15}$ more protons in the low energy state than in the high energy state, giving a net magnetic moment (M_0) parallel to the external magnetic field. Let's call these *excess protons*.

10 Magnetic field strength is measured in tesla (T), with 1 T = 1 kg/(s²A).

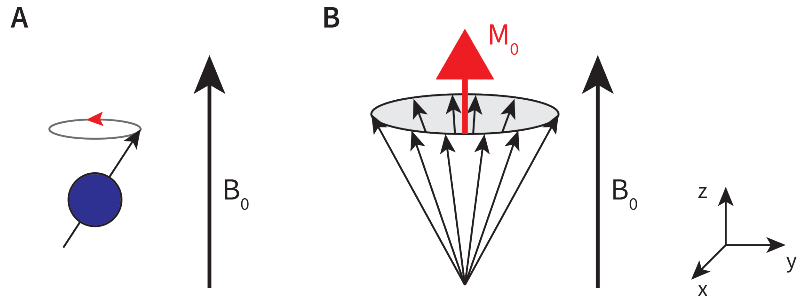


Figure 13.11. Alignment of protons with the external magnetic field. A, precession movement of a single proton. B, the overall precession movements of all protons cancel out on the x and y axes, resulting in a net magnetization moment M0 parallel to the z axis.

In addition to the separations of protons into high and low energy states, the external field also induces a movement of *precession*, which is similar to the movement of a spinning top (Figure 13.11A).

Protons precess with an angular frequency ω proportional to the field strength. This is called the *Larmor frequency*:

$$\omega = \gamma B_0 \tag{13.2}$$

where ω is the Larmor frequency, γ is a quantity characteristic of every nucleus (gyromagnetic ratio) and B_0 is the external field strength.

In an external field of 1 T, the Larmor frequency for ^1H is ~42.6 MHz, at 2 T it will be ~85.2 MHz, and so on. These are all frequencies in the domain of radio waves (see the Photobiology chapter if you need a reminder on the domains of EM radiation).

To visualize the equilibrium state better, let us look at Figure 13.11B. We define a reference frame where the external magnetic field is aligned with the *z* axis. At equilibrium, the net magnetization moment will be aligned with this axis. There is no component of the net magnetic moment vector on the *x* or *y* axis, as the precession movements of all the spinning protons cancel out their components on the *x* and *y* axis.

4.2. Nuclear magnetic resonance. Basis of MRI

We have described what happens at equilibrium when protons are placed in an external magnetic field (Figure 13.10). Overall, due to the slightly higher population of protons in the lower energy state, the protons will have a net magnetic moment M_0 that is aligned in the direction of the external field. However, M_0 is tiny compared to the external field B_0 and can't be measured directly. What we can do instead is perturb the established equilibrium. In brief, a radiofrequency EM radiation pulse is applied to the protons, thereby altering their state. Finally, the radiofrequency pulse is stopped and the protons are left to relax back to the equilibrium state, a process described by two time constants, T1 and T2. The detailed description of the process is provided below.

The Larmor frequency also has another physical meaning, it is the frequency corresponding to the energetic difference between the parallel and antiparallel states (shown as ΔE in Figure 13.10) or the *resonance frequency*. If a radiofrequency (RF) pulse is applied to our proton volume in the *xy* plane at the Larmor frequency, protons in the lower energy state will absorb this energy and switch to the higher energy state, leading to a change in the net magnetization vector. This is called *excitation*. Excitation with an RF pulse will have two effects on the protons:

- it converts low energy to high energy protons;
- it provides *resonance*, causing the precession movement of the protons to be in phase (synchronized);

Following excitation, the magnetization vector will be rotated from the initial direction by a certain angle that we call the *flip angle*, which depends on the time that the pulse is applied, and thus on the total energy provided to the protons. Commonly used flip angles in MRI are 90° and 180°.

The description of an entire NMR experiment is shown in Figure 13.12. Overall, this will result in determining two time constants, T1 and T2 that depend on the environment in which protons are situated. Let's see what these mean.

Let's consider what happens when the flip angle is 90°. In order to obtain this angle, the energy provided must exactly flip half of the excess protons that are in the low energy state to the high energy state. Thus, applying a 90° RF pulse will reduce the longitudinal (*z* axis) component of M_0 to zero, and will give a maximum transverse component (magnetization in the *xy* plane) of the net magnetization vector, as the movements of the protons are now in phase (Figure 13.12A). Once the RF pulse has been turned off, protons in the excited state will naturally return to their equilibrium state through a process called *relaxation*. During relaxation from a 90° pulse, a sinusoidal electric signal can be recorded by the MRI machine called the *free induction decay*

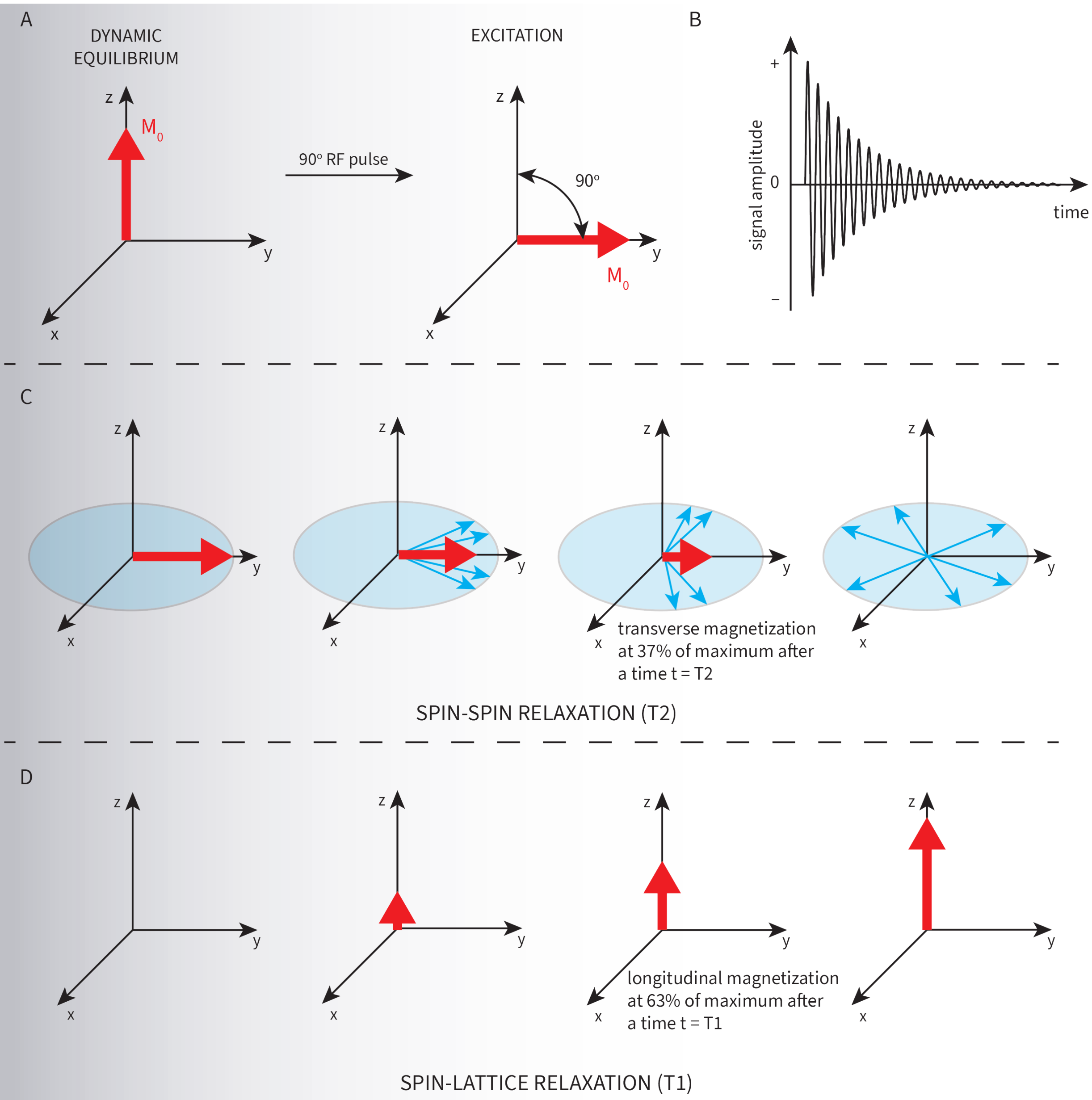


Figure 13.12. Nuclear magnetic resonance and signal generation in MRI. A, excitation using a 90° RF pulse. B, the FID curve shows the recorded MRI signal as transverse magnetization decays to zero. C, the decay over time of transverse relaxation to zero is called spin-spin relaxation and is described by a time constant called T2. Blue arrows show dephasing of the proton spins. D, the return over time of the longitudinal relaxation to maximum is called spin-lattice relaxation and is described by a time constant called T1.

(FID). The FID signal (Figure 13.12B) follows an exponential decay, during which the spins of the protons stop rotating in phase, and return to an overall zero transverse magnetization. This can be mathematically described by:

$$M_{xy}(t) = M_0 e^{-\frac{t}{T2}} \qquad (13.3)$$

where $M_{xy}(t)$ is the transverse magnetization at time *t*, M_0 is transverse magnetization at $t = 0$ and T2 is the time constant of the exponential decay curve.

The process (Figure 13.12C) is called *spin-spin relaxation* (*T2 relaxation*) and it is characterized by the time in which transverse magnetization decays to 37% (1/e) of the maximum, which is called the *transverse relaxation time* or *T2*.

Relaxation also leads to recovery of the longitudinal component (Figure 13.12D), as excited protons will flip back to the lower energy state. This is called *spin-lattice relaxation* (*T1 relaxation*). We will skip a lot of the technical details here and simply say that, even though we cannot detect the component of M_0 on the *z* axis directly, this can be done indirectly, by performing several successive 90° RF pulses.

We characterize the spin-lattice relaxation by the time that it takes for 63% (1 – 1/e) of the longitudinal magnetization to be recovered, which we call *longitudinal relaxation time* or *T1*.

Table 13.3. Approximate T1 and T2 times for different types of tissue[11]. CSF = cerebrospinal fluid.

Tissue type	T1 for B_0 = 1.5 T (ms)	T2 (ms)
Fat	260	80
Liver	500	40
Muscle	870	45
White matter	780	90
Gray matter	900	100
CSF	2400	160

4.3. The MRI image

As in CT, we can divide the volume of tissue that needs to be imaged into voxels. Obtaining a signal in MRI thus means determining the T1 and T2 time constants from the respective voxel and converting these into the intensity of a pixel in the MRI image, which is, as in CT, on a grayscale.

In order to be able to divide the imaged tissue into slices, a gradient is applied to the magnetic field imposed on the patient's body. Thus, protons in different planes (axial slices) will be exposed to a slightly different B_0, and thus resonate at slightly different Larmor frequencies, aiding in the detection of the FID signal. Once a signal is acquired from a certain plane, a process called Fourier transform is used to separate the FID signals of all the image elements in the area of interest, resulting in T1 and T2 values for each individual pixel.

Both the T1 and T2 depend on the chemical environment in which protons are situated. A comparison of approximate T1 and T2 times for different tissues is given in Table 13.3. Note that T1 also depends on the strength of the external magnetic field, B_0, while T2 is unaffected by B_0. As a general rule, T1 values are much longer than T2 values.

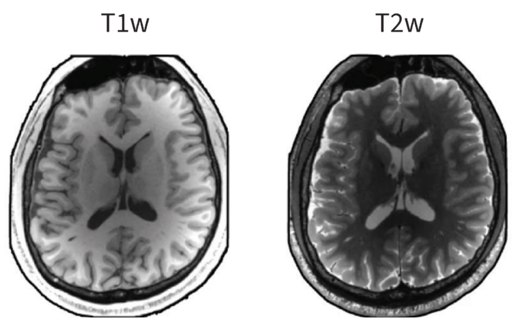


Figure 13.13. MRI images[12]. Comparison of a T1 and a T2 weighted image. Acquisition was performed using a magnetic field of B_0 = 3 T.

Thus, the T1, T2 times and the proton density are characteristic of each type of tissue and can be translated into an MRI image. To aid in visualization, the generated image can be "weighted" in order to produce contrast mainly on the differences of a certain parameter (Figure 13.13).

Figure 13.13 shows an example of two modes of weighting (several more modes exist): T1 and T2 weighting. The T1 weighted image produces contrast based on the T1 characteristics of the tissue: the shorter the T1, the brighter the signal. In a T1 image, fat appears as white, while cerebrospinal fluid (CSF) appears as dark. A T2 weighted image produces contrast based on the T2 characteristics of the tissue: the longer the T2, the brighter the signal. In a T2 image, both CSF and fat are bright.

As in radiography or CT, *contrast agents* can be used in MRI in order to improve contrast in the MRI image. These are usually compounds of gadolinium, a paramagnetic rare earth element. These are generally better tolerated than the iodine-based contrast agents in CT. However, their use is not without risk. For example, patients with kidney failure can develop a potentially life-threatening condition called nephrogenic systemic fibrosis following administration of gadolinium-based compounds.

4.4. Advantages and disadvantages of MRI. Comparison with CT

Let us conclude this section with an overall discussion of the advantages that MRI brings. By far the major advantage compared to CT is the fact that **MRI does not expose the patient to ionizing radiation**. Additionally, contrast in MRI can be better for soft tissue than in CT.

There are, however risks involved with MRI, which is why we did not classify it as a non-invasive technique, but a minimally invasive one. We already listed above the risk of using contrast agents. Additionally, as MRI uses a strong magnetic field, any kind of paramagnetic or ferromagnetic material present in the patient's body is a potential danger. Examples can be: inserted pacemakers, metal implants or even some tattoos. A thorough evaluation has to be performed before the MRI to determine if the patient is suitable for the procedure or not.

Compared to CT, MRI is also much slower, making it unsuitable for emergency situations. Also, the MRI tube can be claustrophobic, especially

11 Data according to Bushberg, J. T., Seibert, J. A., Leidholdt, E. M., & Boone, J. M. (2021). *The Essential Physics of Medical Imaging*: Wolters Kluwer.

12 Image available under a Creative Commons license (*http://creativecommons.org/licenses/by/4.0/.*) from Chen, X., Qu, L., Xie, Y., Ahmad, S., & Yap, P.-T. (2023). A paired dataset of T1- and T2-weighted MRI at 3 Tesla and 7 Tesla. *Scientific Data*, 10(1), 489. doi:*10.1038/s41597-023-02400-y*

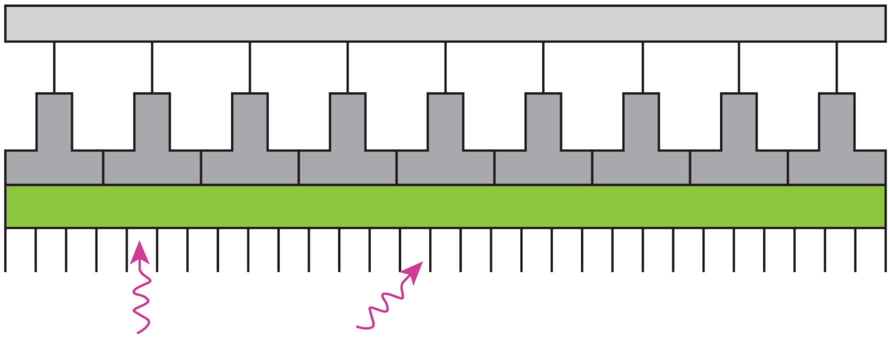


Figure 13.14. A gamma camera. Description is provided in the text.

given the long acquisition time (usually 30 – 90 min).

5. NUCLEAR MEDICINE

5.1. General notions

We call nuclear medicine all medical techniques that involve the use of radioactive isotopes. These techniques insert biologically active compounds containing radioactive isotopes into the body of the patient, which then distribute themselves differently depending on their biological and chemical properties. Finally, an image can be obtained by observing the ionizing radiation emitted by the patient's body as these radioisotopes decay.

Virtually all radioactive isotopes used for imaging are produced artificially. This can be done in either (moderately) small devices called cyclotrons, that are placed in the hospital itself, or in nuclear reactors.

A thorough discussion of radioactivity has been provided in the chapter on Radiobiology, so please review those notions before reading this section. Inserting radioactive isotopes into the body of a patient is highly invasive, as it exposes the patient to ionizing radiation internally, which, as we saw, is generally more harmful than an external exposure. It should be obvious that strong justification is needed for subjecting a patient to such imaging procedures (for example, detection of cancer metastases).

5.2. The gamma camera

The basis of imaging in all nuclear medicine techniques is a device called a *scintillation camera* or *gamma camera* that allows the detection of high energy photons emitted from the patient's body. Note that charged ionizing radiation emitted by radioisotopes present in the body cannot be detected due to its lower penetrative power, and, thus, quick absorption by tissue.

The image acquisition problem is slightly different in nuclear medicine compared to radiography. While in radiography X-rays are directed at the patient in the form of a conical beam, in the case of nuclear medicine techniques, radiation is emitted equally in all directions from the source. Thus, the gamma camera is built in such a way as to both detect the energy of the incoming photons, but also to only detect photons that follow certain trajectories and ignore the rest. This is done through the use of *collimators* (Figure 13.14).

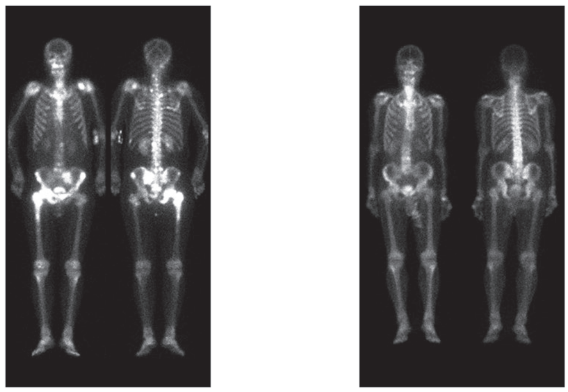


Figure 13.15. Whole-body bone scan[13] performed using ^{99m}Tc of a patient with bone metastasis (a) and one without bone metastasis (b). In both (a) and (b), the left image is an anterior view, and the right a posterior view. Technetium 99m-methyl diphosphonate was injected into patients' veins and image acquisition was performed 4h later. The compound accumulates in areas of active bone formation allowing identification of metastatic regions.

13 Image available under a Creative Commons license (*https://creativecommons.org/licenses/by/4.0/*) from Yu, P.-N., Lai, Y.-C., Chen, Y.-Y., & Cheng, D.-C. (2023). Skeleton Segmentation on Bone Scintigraphy for BSI Computation. *Diagnostics*, 13(13), 2302. Retrieved from *https://www.mdpi.com/2075-4418/13/13/2302*

The collimators of the gamma camera are a grid of lead plates that absorb incoming γ photons that are not parallel to the orientation of the collimator's plates. Photons that pass the collimator then encounter a *scintillation crystal*. The scintillation crystal is made out of thallium-activated sodium iodide, NaI(Tl). When it is hit by gamma photons, it *scintillates*, emitting visible and UV photons. These UV and VIS photons are then detected by *photomultiplier tubes*, that convert the incoming light signal into an electrical signal that can then be processed and analyzed.

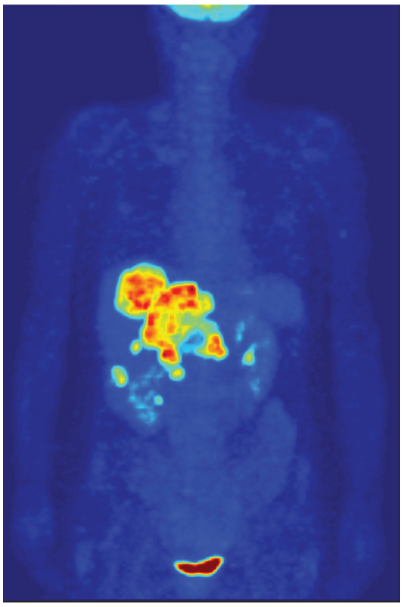


Figure 13.17. PET image showing liver metastases of a colorectal tumor[14] performed using ^{18}F-FDG (anterior view). Normal accumulation of 18F-FDG occurs in the heart, bladder and kidneys and brain.

5.3. Scintigraphy

Scintigraphy or *gamma scan* is a technique that detects γ radiation emitted as consequence of radioisotopes inserted into a patient's body. The image provided by scintigraphy is bidimensional (2D), unlike the images provided by PET or SPECT (see below), which provide 3D images. Thus, one could compare it to a radiographical image, as it contains no depth information.

Examples of isotopes used in scintigraphy are ^{131}I (for imaging of the thyroid), ^{99m}Tc (an isotope of technetium in an excited, metastable state, used for imaging of lung, brain, heart, skeleton, etc.), ^{201}Tl (for imaging of the heart), etc. A sample scintigraphic image is shown in Figure 13.15.

5.4. Positron Emission Tomography (PET)

Positron emission tomography (PET) employs radioisotopes that produce β^+ radiation (positrons). The PET images can be considered to be somehow analogous to those in CT (where, instead, the radiation source is external), as cross-sectional images can be obtained. In practice, due to the similar construction, PET systems can be coupled to CT systems into PET/CT machines, with the bed passing through both the PET and CT detectors.

Positrons cannot be directly detected using

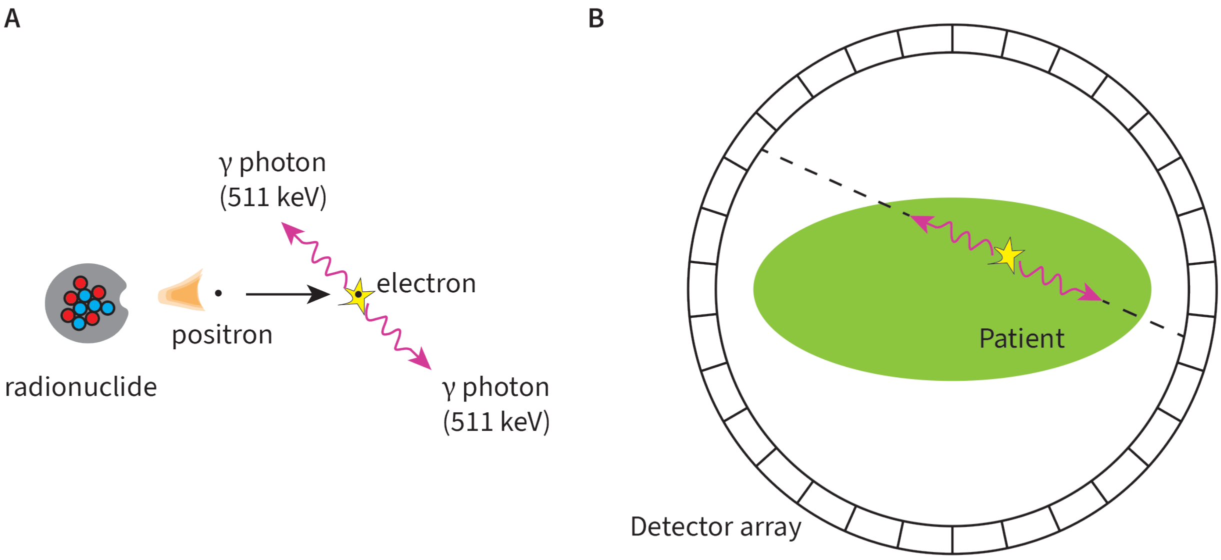


Figure 13.16. Principle of PET. A, annihilation of a positron produced by β^+ decay with an electron from the tissue results in the emission of two γ photons in opposite directions. B, an annihilation event in the body of a patient will release two γ photons that will reach the detector array in opposite positions at approximately the same time. The yellow star represents the annihilation event in both panels.

14 Public domain image by Jens Maus, retrived from *https://en.wikipedia.org/wiki/File:PET-MIPS-anim.gif*

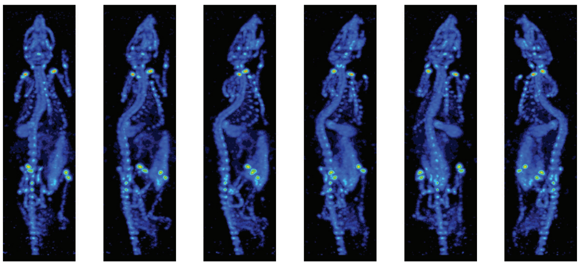


Figure 13.18. SPECT whole-body bone scan[15] of a mouse performed using ^{99m}Tc.

cameras placed outside the body, as these quickly meet an electron of the tissue, triggering an annihilation event. What we can detect, instead, are the consequences of the annihilation event, which converts the masses of the positron and electron into two high energy (511 keV each) γ photons that are emitted in opposite directions (at 180° to each other). Thus, detectors in PET (Figure 13.16) are built to identify annihilation coincidence events (photons that reach opposite sides of the detector at roughly the same time).

One of the most commonly used compounds in PET is ^{18}F-fluorodeoxyglucose (^{18}F-FDG). This compound results from replacing the –OH group in position C2 of glucose with a β^+ emitter, ^{18}F. ^{18}F-FDG is a glucose analogue that cannot be fully metabolized and is retained in the cells. Tumors show up in the PET image as "hot spots", as the tumor glucose metabolism is higher than that of normal tissue (Figure 13.17).

The major advantages of PET are the high contrast and the high spatial resolution of the obtained images. Main disadvantages of PET are its high cost and the short half-life of the used isotopes. These disadvantages are improved in the related technique called SPECT, though at the cost of image quality.

5.5. Single-Photon Emission Computed Tomography (SPECT)

SPECT is similar to scintigraphy, in that radioisotopes that emit γ photons are used. In fact, the same radioisotopes used in scintigraphy can also be used in SPECT. The advantage of these radioisotopes compared to those used in PET is that they are more easily produced and have higher half-lives, thus reducing their cost. Unlike the 2D images obtained in scintigraphy, the SPECT machine is able to provide 3D images (Figure 13.18) as well as images of cross-sections. Thus, in order to obtain the SPECT image, the gamma camera is rotated around the patient's bed.

The main disadvantages of SPECT compared to PET are the lower spatial resolution and the lower sensitivity due to the use of lead collimators compared to coincidence detection.

6. ULTRASONOGRAPHY

6.1. General principles

Ultrasounds are acoustic waves with frequencies above the human hearing range (> 20 kHz). In medicine, ultrasounds with frequencies of ~1 – 20 MHz can be used both in imaging and for their destructive properties (for example, in breaking kidney stones). The latter use of ultrasounds is described in a different chapter; we will refer in this section only to their use in imaging, called *ultrasonography* or *ecography*. As ultrasonography is also described in the practical activities, this section will only give a brief overview of ecographic techniques.

Ultrasounds are produced and also detected through the use of *piezoelectric crystals*. These are crystals such as lead zirconate titanate ($Pb[Zr_xTi_{1-x}]O_3$, with $0 \le x \le 1$) that manifest the following properties:

- when subjected to an external mechanical stress, they generate a potential difference (*piezoelectric effect*);
- when subjected to a potential difference, they vibrate, producing mechanical waves (*reverse piezoelectric effect*).

Thus, production of ultrasounds is performed through the reverse piezoelectric effect, while detection occurs through the piezoelectric effect.

A key principle in ultrasonography is *acoustic impedance* (Z), which was described in the previous chapter on the Biophysics of hearing. Acoustic

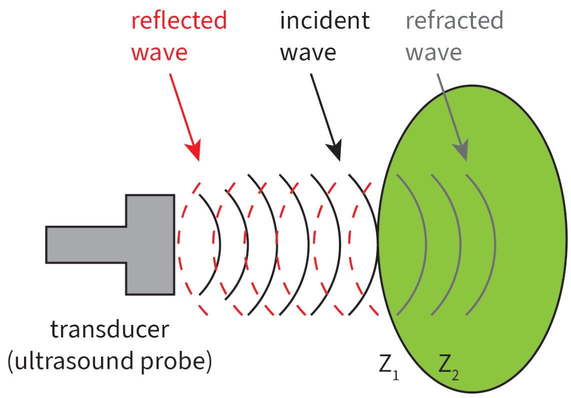


Figure 13.19. General principle of ultrasonography. A transducer directs ultrasound into the tissue. When a different tissue (with $Z_2 \neq Z_1$) is encountered, part of the ultrasounds will be reflected back to the transducer, providing an ecographic signal.

15 Image available under a Creative Commons license (*https://creativecommons.org/licenses/by-sa/3.0/deed.en*) from Christian Lackas (*https://commons.wikimedia.org/wiki/File:Mouse02-spect.gif*).

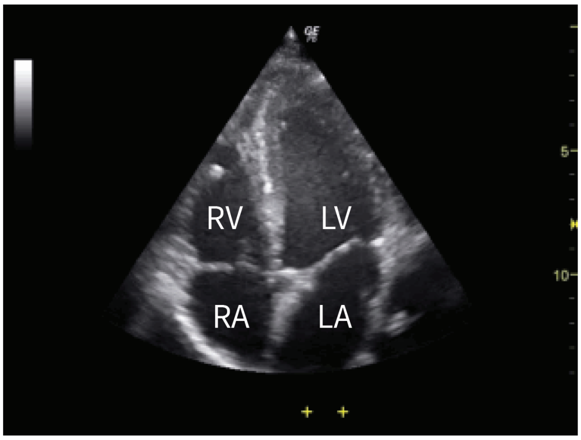


Figure 13.20. Ecographic image of the heart showing an apical view[16]. The heart chambers are annotated for ease of reading. LV = left ventricle, RV = right ventricle, LA = left atrium, RA = right atrium.

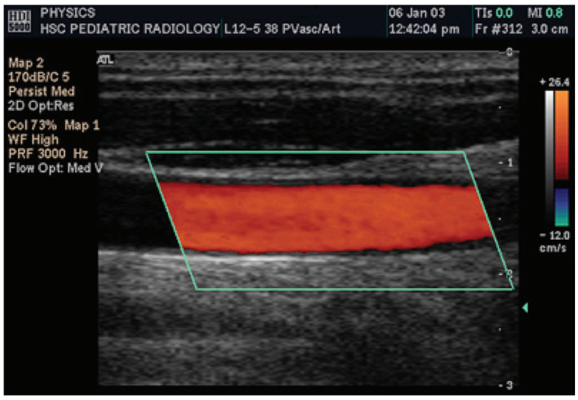


Figure 13.21. Color Doppler image[17] of the common carotid artery. Blood flow towards the transducer is shown in shades of red, blood flow away from the transducer is shown in shades of blue. The color scale on the right shows the correspondence between color and blood flow velocity.

impedance is measured in Rayl (1 Rayl = 1 kg·s/m²). Acoustic impedance describes the opposition of a certain material to the passage of acoustic waves and is calculated as:

$$Z = \rho c \tag{13.4}$$

where ρ is the density of the material and c is the speed of sound in the respective material.

When acoustic waves such as ultrasounds meet the boundary between two different acoustic media (media with different acoustic impedance), a part of the acoustic waves will return to the first medium (reflection), while another part will continue their passage into the second medium (refraction). These are phenomena that are analogous to the behavior of an EM wave that travels between media of different refractive indices.

Ultrasonography relies on directing ultrasounds into the body of the patient and then detecting their reflection back to the transducer (probe). Thus, the *echo* generated by the ultrasounds is detected, hence the alternate name of *ecography*. The general principle is shown in Figure 13.19.

Examples of acoustic impedances for different materials were given in the previous chapter on the Biophysics of hearing in Table 11.1. As a general rule, the higher the difference between the impedances of two media is, the more reflection will occur, as previously shown in equation (11.10). In extreme cases, such as when there is a boundary between tissue and air or bone, most of the ultrasounds will be reflected and virtually none will be transmitted. Thus, both air and bone will create shadows on the ultrasound image behind which further imaging is not possible.

The high reflection between air and skin is also the reason why *ultrasound gel* is used when performing ultrasonography: not using gel would transmit the ultrasounds into the air first, and practically all of them would be reflected back from the air – skin boundary. The acoustic impedance of the ultrasound gel is precisely formulated to be close to the acoustic impedance of the skin, and thus to minimize reflections at the skin's surface.

6.2. Modes of ultrasonography

The most commonly used mode of ultrasonography is the *B mode*, where B stands for brightness. In this mode, the **intensity of the ultrasound echo is converted into brightness** on a grayscale, while the **time needed for detecting the reflected echo is translated into distances**. 2D images can be obtained showing anatomical structures of interest (Figure 13.20).

Other modes of ultrasonography exist. We will only show one more, *color Doppler* (Figure 13.21), which involves the use of the *Doppler effect* to visualize the velocity of moving tissues (blood). The Doppler effect represents the shift in the apparent frequency of a wave depending on the movement of the source relative to the observer. This can be easily understood if you think about how you hear the sirens of an ambulance: when the ambulance is travelling towards you the sounds are high-pitched (high frequency), while when it is moving away from you, the sounds become low-pitch (low

16 Image by Fruehaufsteher2, available under a Creative Commons license (*https://creativecommons.org/licenses/by-sa/3.0/deed.en*), retrieved from *https://en.wikipedia.org/wiki/File:Ultrasound_of_human_heart_apical_4-cahmber_view.gif*. Labeled for clarity.
17 Image by Daniel W. Rickey, available under a Creative Commons license (*https://creativecommons.org/licenses/by-sa/2.5/deed.en*), retrieved from *https://commons.wikimedia.org/wiki/File:ColourDopplerA.jpg*

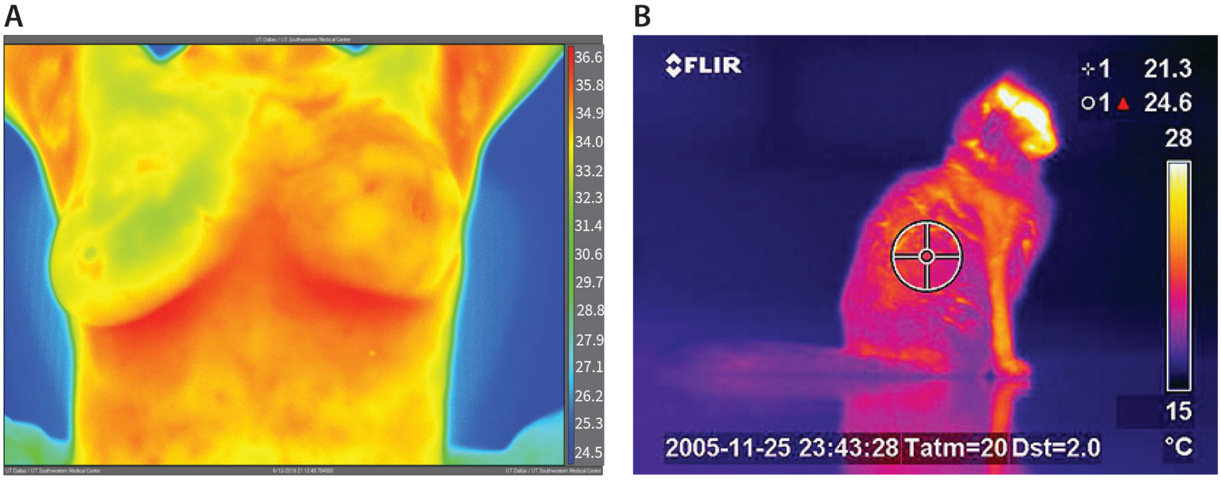


Figure 13.22. Thermograms. A, thermogram of a patient's breasts[18] with normal right breast and left breast affected by cancer. Besides the obvious deformed aspect of the left breast, clear temperature asymmetry is observed between the breasts: temperatures are increased in the left breast compared to the right one, on average, by 1.2 °C. B, thermogram of a cat.[19] In both panels, the color scale on the right can be used to assess temperature (going from low temperatures on the bottom to high temperatures on the top).

frequency). Color Doppler exploits the shift in frequency when ultrasounds are reflected from a moving tissue (blood flow) and adds on top of the anatomical image of the B-mode a color scale showing whether blood is moving towards or away from the transducer. Color Doppler allows determination of both blood flow direction and speed.

6.3. Advantages and disadvantages of ultrasonography

Unlike radiography, CT, or the techniques of nuclear medicine, ultrasonography does not direct ionizing radiation into the patient's body. Thus, we consider it a minimally invasive technique. Ultrasonography is a routine procedure of observing pregnancy, as ionizing radiation techniques should not be used on pregnant women, due to the high risk to the fetus.

The main disadvantages of ultrasonography are the reduced image quality compared to other techniques as well as the decrease in resolution at higher depths and the impossibility of visualizing structures behind bones or aerated regions.

7. THERMOGRAPHY

All objects (including living organisms) with a temperature above 0 K emit infrared radiation (IR) which can be captured by using specialized cameras, thus yielding a *thermogram* (Figure 13.22). As no radiation is directed at the subject, this is a completely non-invasive technique.

In medicine, thermography can be used as a complementary technique. For example, it has been used in the detection of breast cancer, as tumors appear as areas of higher temperature due to their increased vascularization and metabolic rate. However, it has not been shown to work as a standalone test in breast cancer screening, according to the US Food and Drug Administration, which only recommends its use for that purpose as a complement to X-ray screening (mammography), and not instead of it.

REFERENCES

Atkins, P. W., De Paula, J., & Keeler, J. (2017). *Atkins' Physical Chemistry*. London: Oxford University Press.

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Bushberg, J. T., Seibert, J. A., Leidholdt, E. M., & Boone, J. M. (2021). *The Essential Physics of Medical Imaging*: Wolters Kluwer.

Chen, X., Qu, L., Xie, Y., Ahmad, S., & Yap, P.-T. (2023). A paired dataset of T1- and T2-weighted

18 Image available under a Creative Commons license (*http://creativecommons.org/licenses/by/4.0/.*), published in: Lozano, A., Hayes, J. C., Compton, L. M., Azarnoosh, J., & Hassanipour, F. (2020). Determining the thermal characteristics of breast cancer based on high-resolution infrared imaging, 3D breast scans, and magnetic resonance imaging. *Scientific Reports, 10*(1), 10105. doi:10.1038/s41598-020-66926-6. The image was modified by relabeling the temperature scale for better readability.

19 Image by Lcamtuf available under a Creative Commons license (*https://creativecommons.org/licenses/by-sa/3.0/deed.en*), retrieved from *https://commons.wikimedia.org/wiki/File:Termografia_kot.jpg*

MRI at 3 Tesla and 7 Tesla. *Scientific Data, 10*(1), 489. doi:10.1038/s41597-023-02400-y

Dance, D. R., Christofides, S., Maidment, A. D. A., McLean, I. D., & Ng, K. H. (2014). *Diagnostic Radiology Physics*. Vienna: INTERNATIONAL ATOMIC ENERGY AGENCY.

Franklin, K., Muir, P., Scott, T., & Yates, P. (2019). *Introduction to Biological Physics for the Health and Life Sciences*: Wiley.

Hussain, S., Mubeen, I., Ullah, N., Shah, S., Khan, B. A., Zahoor, M., . . . Sultan, M. A. (2022). Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review. *Biomed Res Int, 2022*, 5164970. doi:10.1155/2022/5164970

Jones, J. G., Mills, C. N., Mogensen, M. A., & Lee, C. I. (2012). Radiation dose from medical imaging: a primer for emergency physicians. *West J Emerg Med, 13*(2), 202-210. doi:10.5811/westjem.2011.11.6804

Kissane, J., Neutze, J. A., & Singh, H. (2020). Introduction to Radiology Concepts. In J. Kissane, J. A. Neutze, & H. Singh (Eds.), *Radiology Fundamentals: Introduction to Imaging & Technology* (pp. 11-14). Cham: Springer International Publishing.

Lahiri, B. B., Bagavathiappan, S., Jayakumar, T., & Philip, J. (2012). Medical applications of infrared thermography: A review. *Infrared Phys Technol, 55*(4), 221-235. doi:10.1016/j.infrared.2012.03.007

Lee, C. I., & Elmore, J. G. (accessed on April 24, 2024). Radiation-related risks of imaging. In R. F. Connor (Ed.), UpToDate: Wolters Kluwer.

Lin, E. C. (2010). Radiation risk from medical imaging. *Mayo Clin Proc, 85*(12), 1142-1146; quiz 1146. doi:10.4065/mcp.2010.0260

Lozano, A., Hayes, J. C., Compton, L. M., Azarnoosh, J., & Hassanipour, F. (2020). Determining the thermal characteristics of breast cancer based on high-resolution infrared imaging, 3D breast scans, and magnetic resonance imaging. *Scientific Reports, 10*(1), 10105. doi:10.1038/s41598-020-66926-6

Mossman, K. L. (1985). Medical radiodiagnosis and pregnancy: evaluation of options when pregnancy status is uncertain. *Health Phys, 48*(3), 297-301. doi:10.1097/00004032-198503000-00006

Mostafavi, S. M. (2021). *COVID19-CT-Dataset: An Open-Access Chest CT Image Repository of 1000+ Patients with Confirmed COVID-19 Diagnosis*. Retrieved from: https://doi.org/10.7910/DVN/6ACUZJ

National Cancer Institute. Retrieved from www.cancer.gov

Suetens, P. (2009). *Fundamentals of Medical Imaging*: Cambridge University Press.

Thayalan, K., & Ravichandran, R. (2014). *The Physics of Radiology and Imaging*: Jaypee Brothers Medical Publishers Pvt. Limited.

Yu, P.-N., Lai, Y.-C., Chen, Y.-Y., & Cheng, D.-C. (2023). Skeleton Segmentation on Bone Scintigraphy for BSI Computation. *Diagnostics, 13*(13), 2302. Retrieved from https://www.mdpi.com/2075-4418/13/13/2302

CHAPTER 14

PHYSICAL FACTORS IN THERAPY

Prerequisite knowledge

- Electromagnetic waves
- Ionizing radiation
- Radioactive isotopes
- Acoustic waves

Following the diagnosis of a medical condition, appropriate treatment should be administered. A multitude of therapies exist that direct a certain physical factor (light, ionizing radiation, heat, etc.) into the body. Due to the large number of potential therapies, this chapter cannot encompass all potential therapeutic techniques that use physical factors. Instead, we will describe some of the most used techniques and their general principles of action.

1. NON-IONIZING RADIATION

1.1. Phototherapy

The therapeutic exposure of patients to non-ionizing radiation is called *phototherapy*. This can be used in a number of inflammatory dermatologic diseases such as: psoriasis, vitiligo, atopic dermatitis, chronic eczema, etc.

Typically, phototherapy is performed using UV radiation. The phototherapeutic mechanisms of UV radiation reduce inflammation and cell proliferation through:

- induction of apoptosis (programmed cell death);
- local immunosuppression (reduction in the in the local immune response);
- stimulation of melanocyte growth (relevant in the depigmentating disease vitiligo).

Depending on the wavelengths and additional treatment protocols, phototherapy can be classified as:

- **UVB therapy**: historically, broad band UVB radiation covering the entire UVB spectrum (280 – 320 nm) was used for treatment of the entire range of dermatologic conditions previously mentioned. This has largely been replaced with narrow band UVB (NB-UVB) that uses UV radiation of 311 – 312 nm, which was shown to be more effective in treatment while exposing the patients to lower amounts of UV radiation. NB-UVB therapy can also be applied using excimer lasers that produce wavelengths of 308 nm. While ameliorating symptoms, UVB exposure comes with adverse effects such as: erythema, pruritus, blistering, photoaging as well as a theoretical risk of developing skin cancer. However, current studies on patients subjected to NB-UVB therapy have so far not proven any increased risk of skin cancer;
- **UVA therapy**: the UVA range (315 – 400 nm) can be further subdivided into UVA_1 (340 – 400 nm) and UVA_2 (315 – 340 nm). Phototherapy using radiation in the UVA_1 domain was initially introduced as an alternative to UVB therapy, as UVA_1 radiation does not induce erythema. UVA_1 is mainly used in treatment of fibrotic skin diseases. Its main disadvantage compared to UVB is the high exposure times required, which can go up to 1 h;
- **PUVA therapy**: uses photosensitizing molecules called psoralens (Figure 14.1) in conjunction with UVA exposure. Psoralen is lipophilic and thus membrane permeant; after administration (oral or topical), it intercalates between DNA bases. In the absence of UVA exposure, psoralen has no effect. However, when photoactivated by UVA, it forms covalent bonds with DNA base pairs, inducing apoptosis and immunosuppressive effects. PUVA therapy is effective in the treatment of psoriasis, vitiligo and atopic dermatitis, but comes with an increased risk of skin burns and skin cancer.

In addition to the previously mentioned uses of UV radiation, visible light can also be used in phototherapy. *Neonatal jaundice* appears as an yellowish discoloration of the skin and sclera in newborns, caused by an increased level of serum bilirubin. The main line of treatment for neonatal

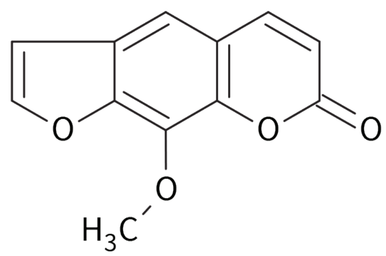


Figure 14.1. 8-methoxypsoralen is a psoralen compound commonly used in PUVA.

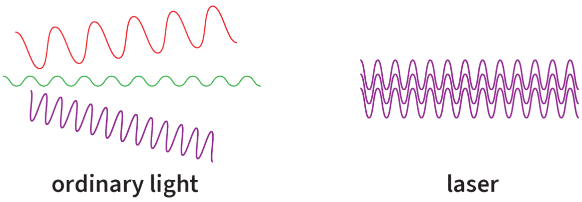


Figure 14.2. Ordinary light sources are polychromatic, incoherent and divergent, whereas laser sources are monochromatic, coherent and collimated.

jaundice is phototherapy using blue or green light. Exposure to light in this range induces bilirubin photoisomerization and conversion to lumirubin, which is polar and easily excreted.

1.2. Laser therapy

Non-ionizing electromagnetic radiation is normally produced as a *polychromatic* and *incoherent* beam (Figure 14.2). Thus, ordinary light sources are characterized by:

- emission of different wavelengths (e.g. a light bulb that emits white light provides EM radiation in the entire visible spectrum) = **polychromatic**;
- the emitted radiation is out of phase (the emitted light waves do not have maxima at the same time or at the same point in space) = **temporal and spatial incoherence**;
- **high divergence**: the emitted light "spreads" in all directions.

If the application of a high intensity light beam is desired, for example in surgery, ordinary light cannot be used. However, light can be used for such a purpose if it is produced through a special process called **l**ight **a**mplification by **s**timulated **e**mission of **r**adiation (**laser**). Laser beams (Figure 14.2) are:

- **monochromatic** (have a single wavelength);
- **spatially and temporally coherent** (have maxima at the same place and at the same time);
- **collimated**, they suffer less divergence as they travel compared to ordinary light beams.

Let's briefly describe the laser production process. We have seen in the previous chapter on Photobiology that absorption of light moves atoms or molecules from the ground state into an excited state (an electron moves from its normal energy level to a higher energy level). In the excited state, atoms or molecules are unstable, and a spontaneous return to the ground state will occur, either by non-radiative processes or by emission of light.

The emission process can also occur non-spontaneously, as *stimulated emission*. The laser device contains an *active laser medium* which can be a crystal, a gas, a liquid, etc. The laser medium is *pumped* in order to bring a high number of atoms or molecules in the excited state. This can be done through directing high powered flashes of light or a high intensity electric current into the laser medium. Pumping is performed in order to bring more atoms or molecules in the excited state than in the ground state – we call this *population inversion*. Spontaneous emission will initiate the laser process: some photons with energies equal to the difference between the excited state and the ground state will be emitted. Stimulated emission can then occur when these photons interact with a molecule in the excited state – this will cause the emission of a second photon, which will have the same energy and direction as the incident photon (Figure 14.3).

The production of lasers is performed in *optical resonators*: the laser medium is contained between two mirrors (Figure 14.4), one of which is a partially reflective mirror (also called a half-silvered mirror). When emitted photons encounter a mirror, they are returned to the laser medium and can provoke further stimulated emission events. Through the partially reflective mirror, some of the photons can escape the resonator and thus form the *laser beam*.

Due to their coherence and collimation, lasers can focus a high amount of energy on a particular spot which can be directed with high precision

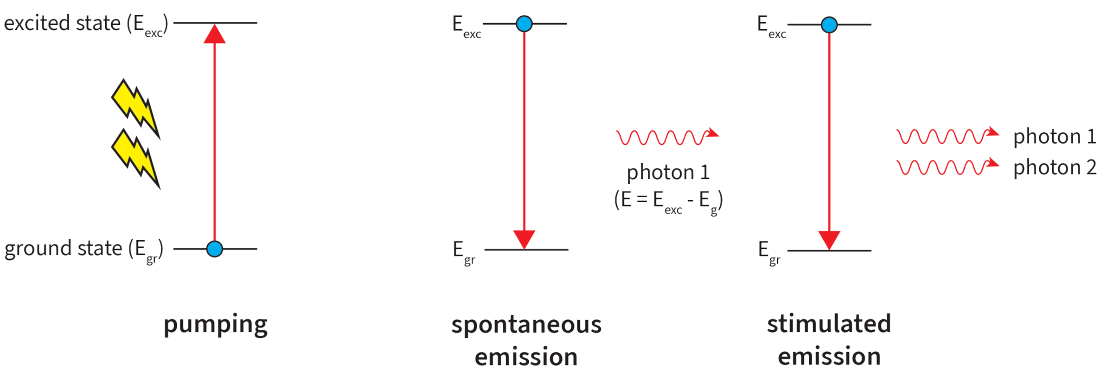


Figure 14.3. The stimulated emission process. Left, atoms or molecules are pumped in order to bring more of them in the excited state than in the ground state. Middle, spontaneous emission occurs, producing a photon (photon 1). Right, photon 1 interacts with another molecule in the excited state, causing stimulated emission of photon 2. E_{exc} and E_{gr} denote the energies of the ground state and excited state, respectively, with $E_{exc} > E_{gr}$. Detailed description is provided in the text.

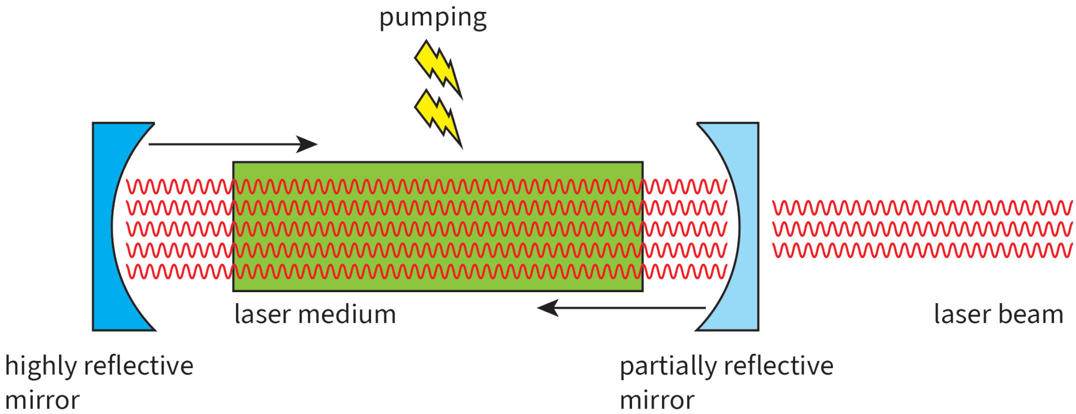


Figure 14.4. An optical resonator produces a laser beam. Arrows indicate photons reflecting back and forth from the surface of the two mirrors. Description is provided in the text.

and exploited for therapeutic use. This represents a high advantage in performing minimally invasive surgery.

We presented in other chapters of this book several types of interaction between electromagnetic radiation and tissue. We can sum these up as: transmission, reflection, scattering and absorption. When lasers are used for a therapeutic purpose, it is the absorption that we are mainly interested in. Depending on the duration of exposure to the laser beam and the power density of the laser, the effects of laser light absorption can be classified as:

- **photochemical effects** (high exposure time, low power density): at low power densities, laser light can be absorbed without significant heating of the tissue. This is used in a treatment technique called *photodynamic therapy*, in which specially designed photoactive compounds called *photosensitizers* are injected into tissue and then irradiated with laser light. The absorption of laser light by the photosensitizer causes excitation, which can lead to formation of reactive oxygen species, followed by cell death through necrosis or apoptosis. Photodynamic therapy is used, among others, in the treatment of autoimmune diseases or cancer;
- **photothermal effects** (medium exposure time, medium power density): the thermal effects are used in surgery for tissue cutting, tissue coagulation or tissue welding. Heating of tissue to more than 60 °C causes irreversible denaturation of proteins. At 90 – 100 °C, the cellular protoplasm is vaporized, followed by carbonization and burning at higher temperatures. An advantage of the photothermal effects is the coagulation of blood vessels, ensuring lower blood loss following laser surgery compared to traditional surgery;
- **photomechanical and photoionizing effects** (short exposure time, high power density): at high beam power densities, non-linear effects start to appear, such as **photoablation**. Photoablation is the removal of portions of tissue without thermal damage to adjacent tissue. It occurs following exposure of tissue to UV lasers with pulse durations of 10 – 100 ns and power densities of $10^7 – 10^{10}$ W/cm^2. Tissue is evaporated by the high intensity laser and, at the same time, ejected from the irradiated area, limiting the damage to the non-irradiated regions. Photoablation allows highly precise tissue remodeling, and is used, for example, in reshaping the cornea for the correction of ametropias in the LASIK (laser-assisted in situ keratomileusis) procedure.

Different types of lasers are used in medicine, each with its own type of laser medium and beam wavelength. We can list: excimer lasers, CO_2 lasers, Ar lasers, infrared lasers, Nd:YAG lasers, etc. The domain of operation extends over the UV, visible and IR domains, with possible wavelengths between 200 nm and 10 μm. At wavelengths above 300 nm, the laser beam can also be directed through a fiber optics tube, allowing endoscopic use.

The detailed description of laser therapies in medicine would require several separate books. Some examples of laser use are in:

- **ophthalmology**: the previously mentioned LASIK surgery, removal of the crystalline lens in cataract surgery;
- **dermatology**: for the treatment of skin diseases, removal of tattoos or skin blemishes, epilation, etc;
- **oncology**: for the removal of tumors through photoablation, heating or via the excitation of a photosensitizer;
- **dentistry**: for tooth or gum reshaping;
- **gynecology**: for the removal of cervical intraepithelial neoplasia (abnormal growths on the surface of the cervix that can evolve into cervical cancer);
- **urology**: removal of ureteral stones (*lithotripsy*);
- **cardiology**: for laser angioplasty (removal of atheromal plaques).

2. IONIZING RADIATION

2.1. General notions

Ionizing radiation is commonly employed for its destructive effect, usually in the removal of tumors. The procedure is called *radiotherapy*. Most patients that suffer from cancer will likely undergo radiotherapy through the course as part of their treatment, either as standalone treatment or combined with surgery or chemotherapy. In particular, in patients with inoperable tumors, radiotherapy is nearly always used.

We have presented the interaction of ionizing radiation with tissue in the chapter on Radiobiology and we saw there that, through direct and indirect effects, ionizing radiation can severely damage DNA. This is equally valid for both healthy and cancer cells! However, studies have proven that cancer cells are less capable of repairing DNA damage and that more DNA breaks are produced in cancer cells than in normal cells. Still, the main disadvantage of radiotherapy remains the inevitable exposure of healthy tissue to ionizing radiation, which can lead to the development of secondary malignancies.

Thus, a major goal of radiotherapy is to maximize the dose delivered to the tumor cells while minimizing the dose delivered to normal cells.

One of the principles of radiotherapy is *fractionation*: the dose is delivered over many treatment sessions, typically as daily fractions of 1.5 – 3 Gy over several weeks. The goal of fractionation is to take advantage of the lower sensitivity of normal cells to ionizing radiation – in the time between the radiotherapy sessions, normal cells have the chance to repair DNA damage before replication.

Further measures can be taken in order to minimize the dose received by the healthy tissue. These involve:

- accurate detection of the tumor's spread through modern imaging techniques;
- modulating the type, position and intensity of the radiation source(s).

2.2. Classification of radiotherapy

Depending on the position of the radiation source, radiotherapy can be classified as:

- **teletherapy (external beam radiation therapy)**: uses an external source of ionizing radiation;
- **brachytherapy**: a sealed radiation source is internally placed as close as possible to the tumor location (sometimes directly inside the tumor). Sources used are placed in such a way that most of the radiation will be absorbed by the patient's body. However, surrounding persons might still be exposed to low doses of ionizing radiation. Thus, patients are advised to maintain distance from pregnant women and children;
- **unsealed source radiotherapy**: a substance containing radioactive isotopes is administered to the patient orally or via injection and the respective substance localizes itself to the site of the tumor through biologically specific mechanisms. In unsealed source radiotherapy, the patient's body will become a source of ionizing radiation as radioactive isotopes may be excreted through sweat or urine or be present in blood and saliva. Thus, adequate isolation of the patient has to be considered in order to protect others, including medical personnel, from exposure to ionizing radiation. The most used radioisotope is ^{131}I, in the form of sodium iodide ($Na^{131}I$). ^{131}I is a β^- emitter and is employed for treatment of hyperthyroidism and thyroid cancer, due to its accumulation in the thyroid.

2.3. Examples of radiotherapy

Both low and high linear energy transfer (LET) radiation is used in radiotherapy. Low LET radiation is represented by EM radiation (X-ray or γ photons). High LET radiation is represented by charged particles (electron, proton or heavy ion beams). In teletherapy, using charged particles such as protons or heavy ions is a more recent development that has the advantage of being able to treat deeper tumors, while reducing the dose that normal tissue is exposed to.

We have seen in the chapter on Radiobiology the shape of the Bragg peak for charged ionizing radiation. In recently developed teletherapy techniques employing proton beams, the energy of the protons can be modulated in order to ensure optimum deposition of radiation across the entire tumor volume. Thus, a "modified proton beam" is

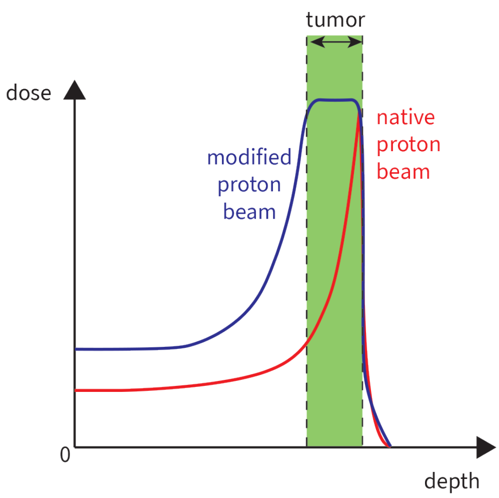


Figure 14.5. The Bragg peak of a modified proton beam is wider, allowing deposition of radiation across the entire tumor volume.

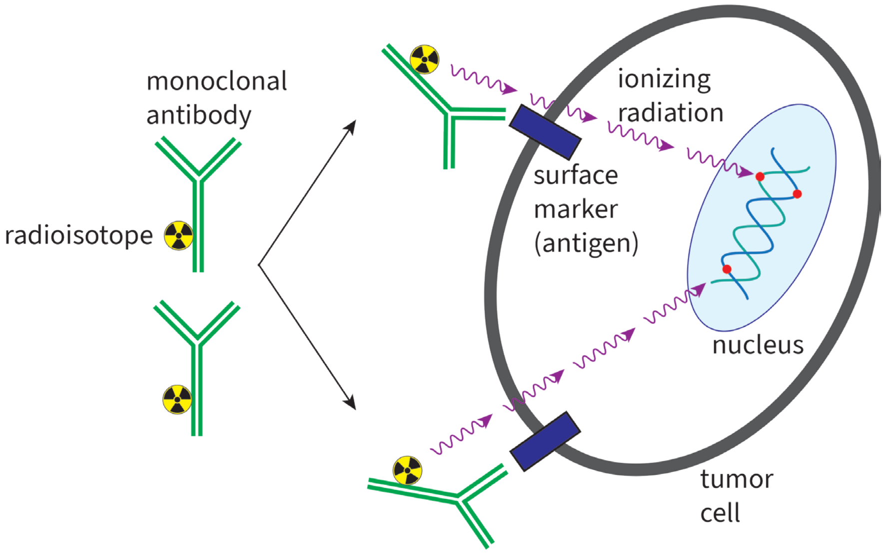


Figure 14.6. Radioimmunotherapy employs antibodies labeled with a radioisotope to target cancer cells. Red circles indicate DNA strand breaks.

obtained by overlapping proton beams of different energies (Figure 14.5).

An emerging field in radiotherapy is that of radioimmunotherapy, which uses antibodies labeled with radioisotopes in order to deliver radiation directly to the tumor site (Figure 14.6). This requires that the tumor expresses a particular cellular surface marker that is not present or accessible in any other cells.

An example of radioimmunotherapy is the treatment of non-Hodgkin's lymphoma (a group of blood cancers) with a mouse monoclonal antibody (ibritumomab) linked to a chelator molecule (tiuxetan) that binds ^{90}Y, a β^- emitter. The functionalized antibody is sold under the name Zevalin. It binds to the B cell marker CD20 and destroys both healthy and malignant B cells. The response rate to the treatment was shown to be ~80%.

3. TEMPERATURE

3.1. Thermotherapy

The use of heat in therapy is called *thermotherapy*. Thermotherapy increases blood flow, tissue temperature, metabolism and connective tissue extensibility, thus promoting tissue healing.

Heating in thermotherapy can be provided through different methods. These can be generally classified as:

- conduction (direct contact with a heated object);
- convection (through a heated fluid);
- conversion of other forms of energy into heat, such as from EM radiation, ultrasounds, electric current, etc.

Thermotherapy has effectively been used in the treatment of musculoskeletal disorders, acute muscular pain, menstrual pain. Note that heat therapy is not recommended if inflammation is present. Injuries associated with inflammation should rather be nursed by applying cold to restrict blood flow.

Increased temperature has also been used in the treatment of tumors, in conjunction with other methods of treatment such as chemotherapy. This is called *hyperthermia* and relies on the fact that, at high temperatures, the cellular metabolism and membrane permeability are increased.

An increase in tissue temperature can also be induced by the use of infrared lasers. Choroidal melanoma of small dimensions can be treated by delivering infrared radiation through the pupil and into the tumor via an IR laser. This results in blood vessel coagulation and necrosis of the tumor cells.

3.2. Cryotherapy

The use of cold in treatment is called *cryotherapy*. The effects of cryotherapy on tissue are generally the opposite of those in thermotherapy. Thus, cold causes vasoconstriction, reducing blood flow, it slows down metabolism and reduces inflammation. Cryotherapy is commonly used in the treatment of acute injury and in pain management.

Cryotherapy can be employed for its destructive effects, in *cryosurgery*. This is most commonly used in the removal of skin lesions such as warts, moles or skin tags and uses liquid nitrogen in order to freeze the tissue. Cryosurgery can also be used in the treatment of some cancers, such as prostate cancer.

Although not usually classified as cryotherapy,

we will mention here the role of cold in *controlled hypothermia* as well as in organ transplant. Controlled hypothermia is used clinically in order to reduce the risk of tissue injury following lack of blood flow. Examples are in the case of cardiac arrest or open heart surgery.

In organ transplant, lowering of the temperature reduces metabolism, allowing longer survival of the organs outside the body and thus facilitating transport. Some tissues (sperm, eggs, embryos) can also be frozen in order to allow long term storage, this is called *cryopreservation*.

4. ELECTRICITY

4.1. Effects of electric current on tissue

Before discussing the uses of electricity in therapy, let us first consider the general effects of subjecting tissue to an electric current.

An effect that occurs when passing any electrical current through a conductive medium is heating. The dissipated heat is calculated according to Joule's law as:

$$Q = I^2 \cdot R \cdot t \qquad (14.1)$$

where Q is the dissipated heat, I is the intensity of the electric current, R is the resistance of the conductor and t is the time for which current flows through the conductor.

The relation between current intensity, resistance and voltage is described by Ohm's law:

$$I = \frac{U}{R} \qquad (14.2)$$

where I is the intensity of the electric current, R is the resistance and U is the potential difference (called *electric tension* or *voltage*).

Thus, Joule's law can be rewritten as:

$$Q = \frac{U^2}{R} \cdot t \qquad (14.3)$$

We see that the amount of heating produced is proportional to the voltage to the power of two. Damage will be greater the higher the voltage, and consequently the intensity of the current, are.

In order for current to flow, the electric circuit must be closed – the charge carriers must have an uninterrupted path to move from high to low potential. Closing of the circuit can occur through contact with the source or contact with the ground. If the circuit is closed through the human body, the pathway that the electric current will take is that of least resistance. In the body, tissues with a high amount of electrolytes and water (blood vessels, muscles, neurons) are good conductors (have low resistance), while tissues with low amounts of water (bone, fat, skin) are poor conductors (have high resistance). If the circuit is closed through the chest, a high risk exists of affecting the function of the heart (see below). Skin provides the highest protection from electric shock, if it is dry. Wet skin has lowered resistance, allowing more current to flow through the body.

If excessive amounts of current travel through tissue, this can cause electrical burns. Unlike thermal or chemical burns, electrical burns can cause severe subdermal damage which might not be immediately apparent. Death or severe injury resulting from electric shock is called *electrocution*.

We have seen in previous chapters that our cells rely on the existence of electric potential differences in order to survive and communicate. Besides direct physical effects such as heating, applying an external electric current can affect the function of the skeletal muscles, heart or nervous system.

The damage that electric injury can impart depends on several factors, including the type of current, intensity and time of exposure and the pathway that the current takes through the body.

Current can be classified depending on its nature (Figure 14.7) as:

- **direct current (DC):** current travels in a single direction due to constant voltage. The source of

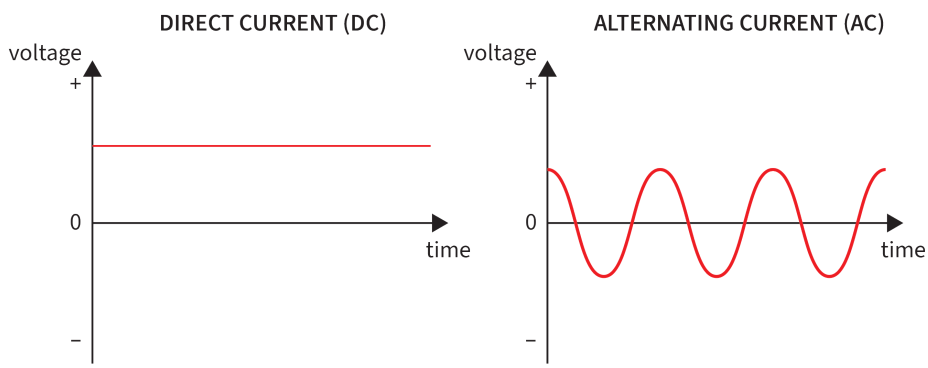


Figure 14.7. Difference between DC and AC. Constant voltage (left) generates direct current, while alternating voltage (right) generates alternating current.

DC has constant polarity (example: a battery);

▸ **alternating current (AC)**: current continuously and periodically reverses direction, due to alternating voltage. The source of AC undergoes a periodical change in polarity: the positive pole turns into negative and the other way around (example: current coming from a wall socket). AC is not characterized only by voltage or intensity, but also by its frequency (how many oscillations occur per unit of time).

The main sources of electric injury under normal circumstances are from household electric appliances and wall sockets. These use AC, at low frequencies of 50 – 60 Hz and voltages of 110 V (in the US, South America and a few other countries) or 230 V (in the rest of the world). Low frequency AC can induce *tetany* in muscles (involuntary muscle contraction). This can cause paralysis of the respiratory muscle or induce ventricular fibrillation, depending on the received current.

Low frequency AC is highly dangerous if the electric conductor is grabbed using the hand. This is because the flexor muscles of the hand are more powerful than the extensor muscles. AC induces repetitive stimulation of both types of muscles. Thus, the victim will grab the electric conductor and be unable to let go, causing extended exposure to the electric current and severe injury. By contrast, DC causes a single muscle contraction that throws the victim away from the source. However, DC is also highly dangerous as it can stop the heart. High frequency AC is less dangerous, as it does not cause tetany.

4.2. Electrotherapy

The use of electricity in medicine is generally called *electrotherapy*. Electrotherapy is used in the management of pain, in increasing joint mobility, promoting tissue repair, relaxation of muscle spasms, etc. Some examples of electrotherapy are:

▸ **Transcutaneous electrical nerve stimulation** (TENS) delivers electric currents to nerves with the purpose of lowering pain. The technique itself is relatively controversial, and there is no universal consensus on which conditions it is most suitable for.

▸ **Diathermy** employs high frequency electric currents in order to heat up deep muscle or joints;

▸ **Electromyostimulation** represents the artificial contraction of muscles using electrical stimuli. This might be used to counteract muscle atrophy in some diseases. Recently, it has been used in sports to increase muscle strength;

▸ **Electroconvulsive therapy** is a technique that applies DC to a patient's head. It is used for the treatment of some psychiatric disorders.

4.3. Other uses of electric currents in medicine

Let us list some uses of electric current in medicine that are not directly classified as electrotherapy:

▸ **Pacemakers** are used in patients with arrythmia to maintain a constant heart rate. Electrodes are placed in the chambers of the heart and the pacemaker is implanted below the clavicle. These monitor the heart rate and, when electrical heart activity is not detected, initiate depolarization via an electrical signal sent to the electrodes;

▸ **Defibrillators** can stop ventricular defibrillation (uncoordinated contraction of the ventricles) by applying an electric current to the heart;

▸ **Electrocautery** uses a small metal probe through which DC flows. The probe is heated up by the Joule effect and can be used to destroy tissue via heating. In electrocautery, electric current does not enter the patient's body;

▸ **Electrosurgery** is different from electrocautery as high frequency AC is used. This current flows through the body of the patient. In the most commonly used variation (monopolar electrosurgery, (Figure 14.8), the current flows from an active electrode (that the surgeon handles) towards a large return electrode placed on the patient's body

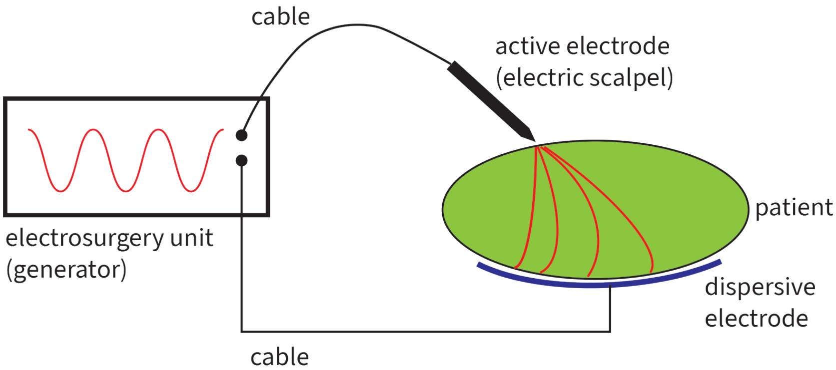


Figure 14.8. Monopolar electrosurgery. The name is slightly misleading, as two electrodes are used, not just one. One is an electrode with a narrow end, providing a high current density (active electrode). The second is a return electrode with a large surface, allowing the closing of the electric circuit without producing heating and burns on the patient's skin. Curved lines suggest flow of the current through the patient's body.

(usually, on the patient's back). Electrosurgery produces destruction of tissue, together with desiccation and coagulation. Compared to using a scalpel, less blood loss will occur in electrosurgery. Risks exist regarding burns or surgical fires which have to be addressed by following adequate safety precautions.

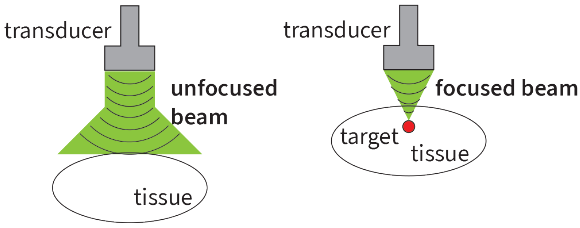


Figure 14.9. Focusing an ultrasound beam allows directing the entire beam's energy at a particular target.

5. ULTRASOUNDS

5.1. Ultrasound effects on tissue

A major use in medicine for ultrasounds is in the imaging of soft tissues (ultrasonography, which was presented in the previous chapter on Medical imaging). However, ultrasounds can also be used for therapeutic purposes.

The physical nature and the production of ultrasounds were discussed in previous chapters. We have seen that ultrasounds directed into the body can reflect or refract when they encounter tissues with different acoustic impedance. Additionally, as ultrasounds travel through tissue, they impart a portion of their energy to the tissue they pass through, resulting in the attenuation of the ultrasound beam. It is this absorption of ultrasounds by tissue that is exploited for therapeutic use. Note that the absorption is different depending on the type of tissue: in general, tissues with higher protein content absorb more ultrasounds.

We can classify the effects that ultrasound has on tissue as:

- thermal effects;
- mechanical effects;
- chemical effect.

Let's discuss them in turn.

The thermal effect of ultrasound absorption is the **heating of the tissue**. At the energies and exposure durations employed in ultrasonography, this heating is minimized. However, when the effect is desired, it can be achieved by applying unfocused ultrasound pulses or by using higher intensity focused beams (Figure 14.9). The effect of high intensity focused ultrasound beams can go beyond mere heating of the tissue, and result in the coagulation of blood vessels and tissues or tissue vaporization.

The mechanical effects of ultrasounds passing through tissue are caused by **cavitation**. When ultrasounds pass through liquid[1], their pressure waves can cause a local drop in the (static) pressure of the liquid below the liquid's vapor pressure. Thus, small gas bubbles are formed through the vaporization of the liquid. When gas bubbles are subjected to high pressures, they can collapse and implode, causing a shock wave that can damage the neighboring cells. Additionally, collapse of the gas bubbles causes local heating of the tissue to extremely high temperatures.

Finally, **chemical effects** occur due to cavitation, as free radicals are formed when the microbubbles rupture.

5.2. Examples of ultrasound therapies

Some examples of the therapeutic uses of ultrasounds are:

- high intensity focused ultrasounds can be used to break up kidney or ureteral stones (*lithotripsy*), through cavitation, allowing their easy elimination. The procedure can be performed by directing the ultrasounds from outside the body (extracorporeal shock wave lithotripsy) or endoscopically;
- in ophthalmology, ultrasounds are used in cataract surgery to break the crystalline lens into small pieces (*phacoemulsification*);
- in dentistry, high intensity ultrasounds are used to break up plaque;
- in sports medicine, low intensity ultrasounds are used to heat tissue, promoting healing and regeneration;
- focused ultrasounds can be used in oncology for the destruction of tumors;
- ultrasonic cavitation can be employed for the destruction of fat cells as an alternative to liposuction;
- using ultrasounds, permeabilization of membranes can be achieved in order to facilitate drug delivery. In particular, ultrasounds can be used for aiding drug absorption through the skin, potentially allowing delivery of molecules with low bioavailability.

REFERENCES

Azadgoli, B., & Baker, R. Y. (2016). Laser applications in surgery. *Ann Transl Med, 4*(23), 452. doi:10.21037/atm.2016.11.51

Baeyens, A., Abrantes, A. M., Ahire, V., Ainsbury,

1 Remember that both the cytoplasm and extracellular medium are liquids.

E. A., Baatout, S., Baselet, B., . . . Wozny, A.-S. (2023). Basic Concepts of Radiation Biology. In S. Baatout (Ed.), *Radiobiology Textbook* (pp. 25-81). Cham: Springer International Publishing.

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Bulat, V., Situm, M., Dediol, I., Ljubicic, I., & Bradic, L. (2011). The mechanisms of action of phototherapy in the treatment of the most common dermatoses. *Coll Antropol, 35 Suppl 2*, 147-151. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/22220423

Durante, M., & Loeffler, J. S. (2010). Charged particles in radiation oncology. *Nat Rev Clin Oncol, 7*(1), 37-43. doi:10.1038/nrclinonc.2009.183

Fish, R. M., & Geddes, L. A. (2009). Conduction of electrical current to and through the human body: a review. *Eplasty, 9*, e44. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19907637

Gallardo-Villagran, M., Leger, D. Y., Liagre, B., & Therrien, B. (2019). Photosensitizers Used in the Photodynamic Therapy of Rheumatoid Arthritis. *Int J Mol Sci, 20*(13). doi:10.3390/ijms20133339

Johnson, M. (2007). Transcutaneous Electrical Nerve Stimulation: Mechanisms, Clinical Application and Evidence. *Rev Pain, 1*(1), 7-11. doi:10.1177/204946370700100103

Knappe, V., Frank, F., & Rohde, E. (2004). Principles of lasers and biophotonic effects. *Photomed Laser Surg, 22*(5), 411-417. doi:10.1089/pho.2004.22.411

Krenitsky, A., Ghamrawi, R. I., & Feldman, S. R. (2020). Phototherapy: a Review and Update of Treatment Options in Dermatology. *Current Dermatology Reports, 9*(1), 10-21. doi:10.1007/s13671-020-00290-6

Lee, S. Y., Fiorentini, G., Szasz, A. M., Szigeti, G., Szasz, A., & Minnaar, C. A. (2020). Quo Vadis Oncological Hyperthermia (2020)? *Front Oncol, 10*, 1690. doi:10.3389/fonc.2020.01690

Lentacker, I., De Cock, I., Deckers, R., De Smedt, S. C., & Moonen, C. T. W. (2014). Understanding ultrasound induced sonoporation: Definitions and underlying mechanisms. *Advanced Drug Delivery Reviews, 72*, 49-64. doi:https://doi.org/10.1016/j.addr.2013.11.008

Miller, D. L., Smith, N. B., Bailey, M. R., Czarnota, G. J., Hynynen, K., Makin, I. R., & Bioeffects Committee of the American Institute of Ultrasound in, M. (2012). Overview of therapeutic ultrasound applications and safety considerations. *J Ultrasound Med, 31*(4), 623-634. doi:10.7863/jum.2012.31.4.623

Nadler, S. F., Weingand, K., & Kruse, R. J. (2004). The physiologic basis and clinical applications of cryotherapy and thermotherapy for the pain practitioner. *Pain Physician, 7*(3), 395-399. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/16858479

Niemz, M. H. (2007). *Laser-Tissue Interactions*. Heidelberg: Springer-Verlag.

Oberli, M. A., Schoellhammer, C. M., Langer, R., & Blankschtein, D. (2014). Ultrasound-enhanced transdermal delivery: recent advances and future challenges. *Ther Deliv, 5*(7), 843-857. doi:10.4155/tde.14.32

Quarato, C. M. I., Lacedonia, D., Salvemini, M., Tuccari, G., Mastrodonato, G., Villani, R., . . . Sperandeo, M. (2023). A Review on Biological Effects of Ultrasounds: Key Messages for Clinicians. *Diagnostics (Basel), 13*(5). doi:10.3390/diagnostics13050855

Rizzieri, D. (2016). Zevalin® (ibritumomab tiuxetan): After more than a decade of treatment experience, what have we learned? *Critical Reviews in Oncology/Hematology, 105*, 5-17. doi:https://doi.org/10.1016/j.critrevonc.2016.07.008

Shields, C. L., & Shields, J. A. (1999). Transpupillary thermotherapy for choroidal melanoma. *Curr Opin Ophthalmol, 10*(3), 197-203. doi:10.1097/00055735-199906000-00008

Sminia, P., Guipaud, O., Viktorsson, K., Ahire, V., Baatout, S., Boterberg, T., . . . Vogin, G. (2023). Clinical Radiobiology for Radiation Oncology. In S. Baatout (Ed.), *Radiobiology Textbook* (pp. 237-309). Cham: Springer International Publishing.

Tempany, C. M., McDannold, N. J., Hynynen, K., & Jolesz, F. A. (2011). Focused ultrasound surgery in oncology: overview and principles. *Radiology, 259*(1), 39-56. doi:10.1148/radiol.11100155

Vieyra-Garcia, P. A., & Wolf, P. (2021). A deep dive into UV-based phototherapy: Mechanisms of action and emerging molecular targets in inflammation and cancer. *Pharmacol Ther, 222*, 107784. doi:10.1016/j.pharmthera.2020.107784

Vilos, G. A., & Rajakumar, C. (2013). Electrosurgical generators and monopolar and bipolar electrosurgery. *J Minim Invasive Gynecol, 20*(3), 279-287. doi:10.1016/j.jmig.2013.02.013

Zaheer, J., Kim, H., Lee, Y. J., Kim, J. S., & Lim, S. M. (2019). Combination Radioimmunotherapy Strategies for Solid Tumors. *Int J Mol Sci, 20*(22). doi:10.3390/ijms20225579

Zemaitis, M. R., Foris, L. A., Lopez, R. A., & Huecker, M. R. (2023). Electrical Injuries. *StatPearls [Internet]*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK448087/

CHAPTER 15

PSYCHOPHYSICS

Prerequisite knowledge

▸ This chapter has advanced topics; we suggest that the students should read this chapter only after they completed previous chapters in Bioelectricity, Biophysics of vision and Biophysics of hearing

1. INTRODUCTION. DEFINITIONS

Psychophysics is the branch of Biophysics which studies the relationships between the *stimuli* from the physical world and the *sensations and perceptions* that they evoke in the human body and mind.

Psychophysics is an overly complex topic which brings together several distinct fields: physics, anatomy, physiology, biophysics, psychology and surprisingly, mathematics (statistics) and information theory (signal processing). It has widespread applications in technology (for instance in all audio-video processing) and in several medical specialties that deal with the senses (audiology, ophthalmology, neurology, psychiatry).

"Classical psychophysics" started as a field in the 1830s – 1860s with the research of E. H. Weber and G. T. Fechner. They discovered an unexpected and consistent logarithmic relationship between intensities of the stimuli and the sensations they evoked (see Figure 15.4 for an example). Since the ~1960s, "modern psychophysics", started to study the complex mechanisms by which perceptions are formed at the neuronal level; this is research in progress and there are still unknown things in this area.

In this introductory chapter for 1st year medical students, we aim to present the core concepts of classical psychophysics, including: relationships between stimuli and sensations, specific vocabulary, sensory modalities, objective measurement methods, information encoding in neuronal systems and their theoretical limitations; applied examples in audiology.

1.1. Specific vocabulary

There are some notions that have a specific meaning in neurology, biological psychology and psychophysics. To avoid any confusion, in this lecture we are using the following notions:

▸ **Stimulus**: a stimulus is a detectable change in the organism's environment. It can be a physical quantity (for instance, intensity of light) or a chemical quantity (for instance, a change in concentration of salt in the mouth). A stimulus is an *objective quantity* (i.e. it exists in the environment; it is something that happens in the environment).

▸ **Sensitivity**: the ability of an organism or organ to detect external stimuli, so that an appropriate reaction can be made, is called sensitivity or excitability.

▸ **Sensation**: the process of gathering information about the world through the detection of a stimulus by a living organism. A sensation arises because of a stimulus that was detected by the organism. A sensation is a *subjective phenomenon* (i.e. it exists in the neuronal system / mind of the person). Different persons can have different sensations at the encounter of the same stimulus.

▸ **Sense**: a sense is a biological system (with anatomical and physiological components) used by an organism for sensation. It should be noted that the word "sense" has a more precise meaning in medical sciences. For instance, five human senses were traditionally identified in human history: sight, hearing, smell, taste, and touch; many more are now recognized (for instance, proprioception – the sense of self movement and body position or chronoception – our ability to sense the passage of time). The actual description of a sense includes both the subjective sensations it evokes and the biological foundations for it (anatomical, physiological, neurological pathways). Thus, a sense (like hearing) consists of several distinct sensory modalities (see below).

▸ **Sensory modality**: one aspect of a stimulus, or of what is separately perceived after a stimulus encounters a sensory organ. For instance "brightness" is the sensory modality that corresponds to luminance (photometric measurement of luminous intensity) of an object. The circles in Figure 15.1 have different luminous emittances and evoke different sensation of brightness. Each human sense has several sensory modalities. For example, the sense of hearing has the following modalities (separately perceived qualities of an

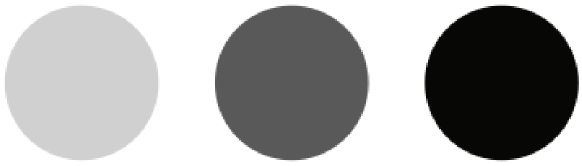


Figure 15.1. Luminous emittance (amount of light emitted by a surface) evokes the sensory modality of brightness. The disks have different emittances and are perceived as having different brightness.

acoustic signal): loudness, pitch, timbre (also known as tone color or tone quality) and localization. Each sensory modality has a different biological mechanism. For instance, pitch is detected by the Fourier analysis done by the basilar membrane in the cochlea, while localization is done in the cortex via a differential analysis of signals from both ears, etc. For all human senses, there are in total about 17 different sensory modalities known.

- **Perception**: the integration of sensory modalities into a coherent, meaningful information. Many sensory modalities and also memory and experience participate in varying degrees to form a meaningful perception. For instance (Figure 15.2), an ambiguous object such as a rope-like object on the floor of a dark room can be perceived as a rope, or as a snake. The mechanisms of perception formation are not yet completely understood and are actively researched. Perception is not limited to one area of the brain: many brain regions are activated when sensory information is perceived from the environment. It is commonly accepted now that the brain integrates multiple sources of information (from each sensory modality) and also from memory and combines them to create a "multimodal perception" – i.e. an internal representation of the outside reality. This research has broader implication for fundamental neurology, psychology and psychiatry, but we will not discuss it further in this introductory chapter.

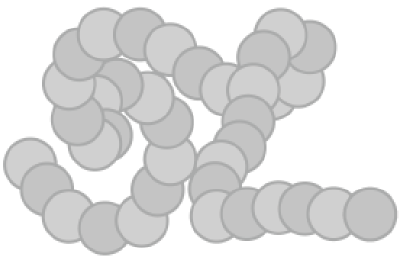


Figure 15.2. A perception is a construct of several sensory modalities and also of memory and experience. The latter are predominant when the amount of information yielded by the senses is not sufficient to form a comprehensive representation.

2. THE OBJECTIVE MEASUREMENT OF HUMAN SENSATIONS

The act of measuring is an objective action, when we are measuring a quantity in the physical world (the weight of an object, the length of a stick, etc). Measuring the intensity of a subjective sensation is a difficult task, since the sensation is dependent on two things: the intensity of the stimulus (an objective fact) and the sensitivity of the person (which is subjective, i.e. depends on the person and the state of the person). For instance, the sensation of the pitch of a sound is dependent on both the frequency of the acoustic signal and also on the aural sensitivity of the person and of the person's emotional state, fatigue, etc. Several ways of measuring were proposed; we present below one that is simple to understand and which has broad practical medical applications: finding thresholds for each sensory modality.

For each person that is tested, a succession of stimuli are produced (for instance, an acoustic signal is generated by a speaker, with a precise amplitude and frequency). In the following examples, "S" denotes the intensity of sensation and "E" the intensity (amount) of the physical stimulus:

- i) We start with the inferior limit of S (S = 0) which corresponds to the minimum amplitude (E) of the stimulus that can still generate a discernible sensation. This is usually called the "inferior threshold of the stimulus" (Figure 15.3).
- ii) We end with a superior limit of S (S_{max}) which corresponds to a stimulus that can no longer be perceived or is so intense that it starts to become painful (E_{max}). This is usually called the "superior threshold of the stimulus".
- iii) The last thing to define is the step of the scale S (the measurement unit). This is not arbitrarily chosen but corresponds to a minimum change in the physical stimulus (ΔE) that it is still noticeable by a human observer. This is also known as the "difference threshold", "just noticeable difference"

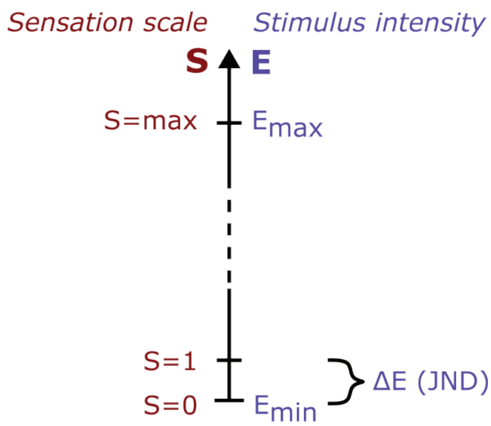


Figure 15.3. The construction of a sensation scale (S), that corresponds to the physical intensity (E) of a stimulus.

(JND) or “least perceptible difference” (Figure 15.3).

This smallest change of E, called *just-noticeable difference* (JND) is the amount by which the stimulus must be changed in order for a difference to be noticeable by the observer, detectable at least half the time (i.e. we test this many times, since the sensitivity of a person differs from time to time). The JND is not the same for all the persons and also the JND is not constant in time for the same person. This makes the building of sensation scales a tedious process.

The newly built scale has a minimum, maximum, and a measurement unit. Scales have been built up for common sensory modalities, and received distinct names (“mels” scale for pitch perception, “son” scale for sound intensity, etc.).

Sensation scales are typically built with the data gathered from a large sample of healthy human subjects. These scales are statistical in nature (they correspond to the average population parameters).

3. THE WEBER-FECHNER LAW

The research done in studying sensations yielded a counterintuitive observation: for the vast majority of the senses, we humans do not perceive well the absolute amount of a stimulus; we are actually more sensitive to the amount of change of the stimulus. We shall give a visual example now, and later an auditory example.

Take a look at the pairs of panels with dots in Figure 15.4 (pair I – II and pair A – B). It is visually obvious that panel II has more points than panel I. But it is not as obvious that panel B also has more points than panel A. Panel II is clearly more crowded than panel I; panel A and B seem to be similarly crowded (as a verification, the number of points is written below each panel).

Weber and Fechner discovered a general numerical relationship between the physical stimulus intensity (E) and the corresponding sensation intensity (S).

Weber found that what we perceive is not the absolute difference ΔE between two intensities E of stimuli (in our visual example in Figure 15.4, E is the number of points in each panel and ΔE = 5 points are added in lower panels). We perceive rather the ratio between this difference and the previous intensity E of the stimulus we were exposed to (ΔE/E).

This ratio (ΔE/E) is called the *relative difference* or *resolution power* of the sense organ.

It was determined experimentally that the resolution power is constant over a broad range of stimuli intensities (ΔE/E = const.) for the same sensory modality. Fechner found that each relative difference is related to a *variation in sensation experienced* (ΔS):

$$\Delta S = k\frac{\Delta E}{E} \qquad (15.1)$$

or:

$$dS = k\frac{dE}{E} \qquad (15.2)$$

where dS and dE are infinitesimal variations of sensation and stimuli intensity (E) and k is a numerical constant which is different for each sensory modality investigated.

Integrating the differential equation (15.2) yields:

$$S = k\ ln\ (E) \qquad (15.3)$$

Equation (15.3) is also known as the

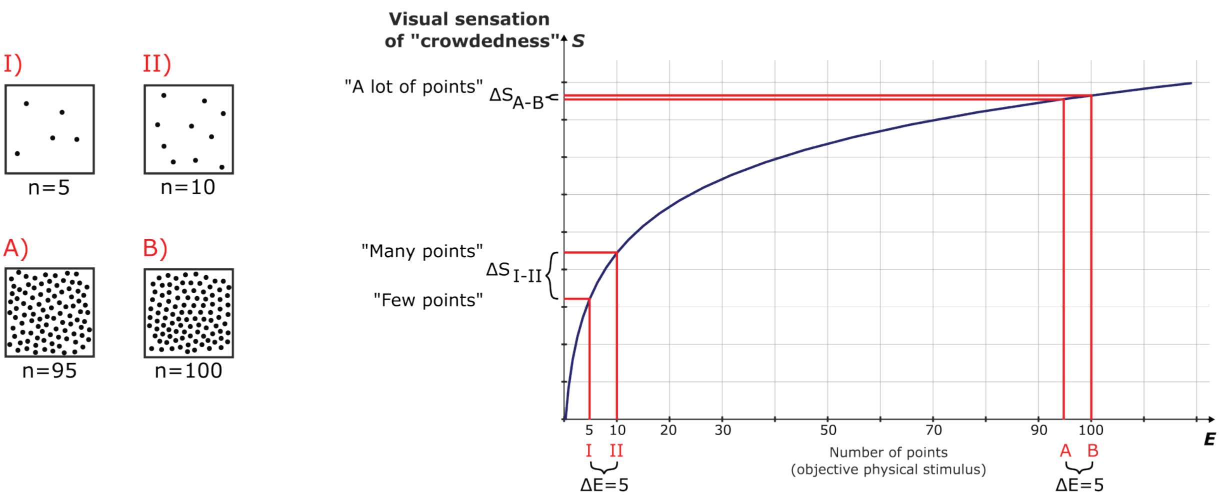


Figure 15.4. Panel II contains 5 more dots than panel I (ΔE = 5 points). Equally, panel B contains 5 more points than panel A (ΔE = 5 points). Yet the changing of the sensation of crowdedness (ΔS) is not equal. In panel II there are clearly more dots than in panel I, but in panels A and B the number of dots appears to be similar.

Weber-Fechner law, which can be summarized as*:* **sensation perceived is logarithmically dependent on the stimulus intensity (for the majority of the senses)**.

4. THE POWER LAW

Extensive research was performed around 1950 and later, done by S. Stevens, on an exhaustive amount of sensory modalities (for instance cold, warmth, tactile pressure feeling, loudness, different tastes, smells, different pain types, etc.). There are sensory modalities where the Weber-Fechner law appears to be ignored (like distance judgement, pain, etc), and in others there is a logarithmic-like relationship. Through a similar laboratory method as devised by Weber and Fechner, but applied to each sensory modality, it has been found that the relationship in these situations can be expressed as a more general law:

$$S = k \cdot E^n \tag{15.4}$$

The above equation is known also as **Stevens' law**, or **the Power law**, n being different for each sensory modality, determined experimentally. See Figure 15.5 for examples of n.

For sensation of perceived brightness of an object vs. actual luminous intensity of that object, $n = 0.5$ (logarithmic-like curve). For the perceived length of an object vs. actual length of that object, $n = 1$ (linear relationship). For the sensation of pain inflicted on the skin by a small electric shock vs. the intensity of the current, $n = 3.5$ (exponential relationship).

As a general conclusion, for the sensory modalities that mostly convey general information about the environment (like brightness, sound intensity, color, etc), the sensitivity curve is logarithmic

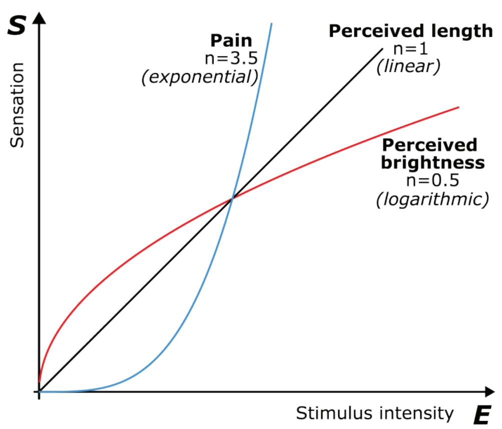


Figure 15.5. The Power law (Stevens' law) of sensation intensity S in relationship with stimulus intensity E. It is a generalization of Weber's law: the relationship is different for each sensory modality.

(Weber-like). This has the advantage of mapping a huge variability of the stimulus (very large E differences) in reduced, comprehensible sensation limits. For the sensory modalities that convey critical information about an organism's survival (different kinds of pain due to various types of cellular or tissue stress), the relationship is exponential (the pain sensation increases dramatically on slight increases of a damaging stimulus). This probably ensures a fast response of the organism in face of a dangerous stimulus (the organism cannot ignore the damage to the tissues).

The values of n were obtained averaging data from very diverse persons. This statistical averaging of Stevens' approach ignores individual differences; a person's sensitivity is only approximately described by these laws (these are statistical, population-wide laws). While these are useful in general practical applications, they are hard to apply to a single individual. These laws were experimentally derived and ignore the neurological underpinning of sensations and perceptions (what is called nowadays the "neuronal correlate of perception"), and the fact that sensation intensity is dependent not only on the stimulus intensity but also on the contextual conditions where perception took place.

5. INFORMATION ENCODING PERFORMED BY THE NERVOUS SYSTEM

A modern approach to find the neuronal basis of sensations and perception (since ~1970) is called Signal Detection Theory (SDT). This is a general framework of communications used in technology and informatics and was found to be also very useful in understanding how the brain perceives the external world. The core of this theory is that for the successful communication between systems, there has to be a system that is able to differentiate between information and noise. In any communication channel (be it an axon of a neuron, a wire between two computers or wireless channel between phones, etc.) at the same time there is an overlap between information bearing patterns (called "stimuli" in biology, "signals" in technology) and the random patterns (called "noise") that inevitably exist in the real world channels. The "noise" can be caused by background stimuli (different from the interesting stimulus), by the spontaneous or erroneous activity of the ion channels that cause electrical potential variations in the membrane, by interfering chemicals, or by other physical factors. The "noise" is random, it might be present or not at the time when information transmission takes place.

When a stimulus reaches a sense, the



Figure 15.6. The neurons in the CNS communicate with each other mainly via electrical changes in their membrane potential.[1] These are called "action potentials", see the Bioelectricity chapter on their genesis and properties. Each individual action potential is shown as a vertical thin line – a "spike" in electrophysiology jargon. Multiple action potentials (spikes) form a "spike train". These spikes (changes of electrical activity) carry the information between neurons.

information it carries must be carried by the nervous system, not the stimulus itself (information in speech is carried by the auditory nerves, not the acoustic signals, etc). Therefore, the main activity of the senses is to perform information encoding and transmission. The information in the stimulus is encoded and then transmitted by our nervous system as groups of on-off electrical signals ("spikes"), see Figure 15.6.

In this introductory chapter we will only focus on the electrical activity of the neurons (neurons use also a host of chemical substances to communicate and modulate the communication; these will be presented later, in the next years of medical school)

But how can the diversity of the information in the external world be represented with only membrane potential variations in neurons? Research done in this field discovered that the human nervous system performs three main types of information encoding:

- frequency encoding;
- spatial encoding;
- temporal encoding.

We'll discuss these three types in turn.

5.1. Frequency encoding

It has been found that the **intensity** of every stimulus is encoded as a **change of the frequency of the neuronal spikes** transmitted to the brain.

As an example: a person in a silent room hears an acoustical signal (Figure 15.7). The sound is perceived by the auditory sense (see the chapter on the Biophysics of hearing). The cochlear (auditory) nerve electrical recording shows a a change in the frequency of spikes that are transmitted by the cochlea during these events. The intensity of the acoustic signal (in this case silence vs. sound) is encoded as the increase of the frequency of the spikes in the auditory nerve.

Important observation: a counterintuitive phenomenon was discovered by the research of the sensory activity. The sensory apparatus is *continuously* sending spikes of signals to the brain (we call these the "spontaneous activity"), even when there is no stimulus in the environment. The main characteristic of this spontaneous activity is that the spikes are randomly distributed while there is no stimulus in the environment (these are a continuous, background "noise" as the Signal Detection Theory calls it). The molecular basis for this activity is understood to be the normal, random thermodynamic fluctuations of the

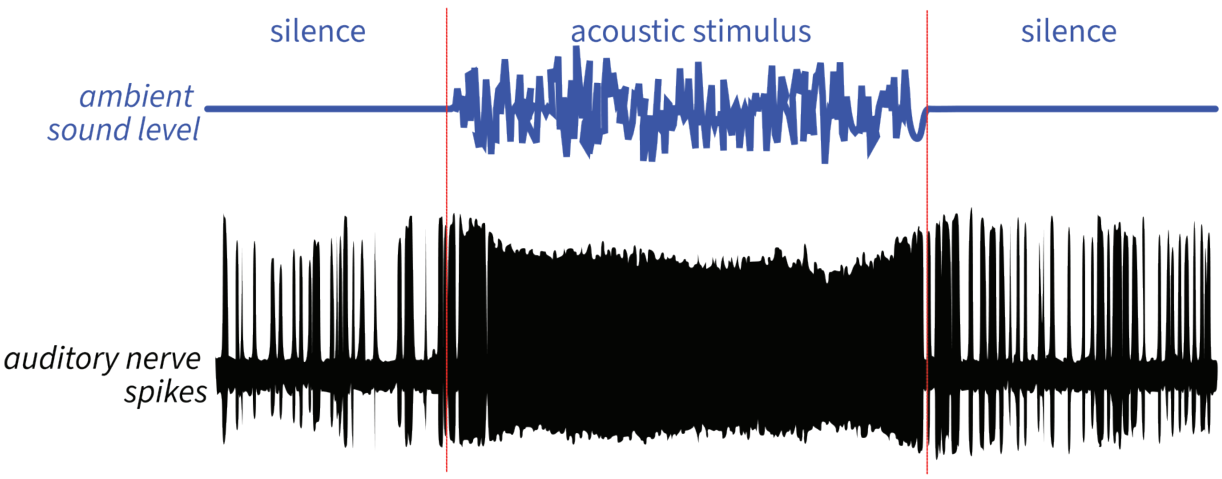


Figure 15.7. Recording of the neuronal activity in the acoustic nerve shows spontaneous, continuous and random activity during the silence periods and an increase in frequency of the electrical activity (many more spikes) while a sound is played.

1 Image is adapted from Stratton, P., Cheung, A., Wiles, J., Kiyatkin, E., Sah, P., & Windels, F. (2012). Action potential waveform variability limits multi-unit separation in freely behaving rats. *PLoS One*, 7(6), e38482. doi:10.1371/journal.pone.0038482, paper available under a Creative Commons license (*https://creativecommons.org/public-domain/cc0/*).

microenvironments near each ionic channel of the membrane. This leads to opening/closing of very few of these channels at random times.

The probability of these events is very low per channel, but given the great number of channels per neuron, it is possible that from time to time, an action potential is triggered by these thermodynamic fluctuations, not by the stimuli. This poses a challenge for the central nervous system: how can it detect when an action potential is from a stimulus (true information) and when it is not (it's a "noise"). Within Signal Detection Theory it was discovered in practice that in such situations, when the transmission of the information is done through a noisy channel, one of the possible approaches is to encode information as not merely the presence and absence of signals, but as a change in the frequency or a change in the patterning of these signals, etc. It seems that the nervous system uses at least these two. When stimulus is absent, the *distribution* of the spikes is truly random (thermodynamic based) and its spectral distribution is flat (i.e. it has a diverse range of frequencies when the Fourier analysis is performed); see section 6 for a brief introduction in Fourier analysis. When a stimulus is present, a pattern of spikes, with a higher frequency is transmitted (in neuroscience jargon, a "train of spikes"), and its frequency distribution is no longer flat; after a Fourier analysis, a clear change in spike frequency can be unambiguously detected.

Let's consider another example: light intensity is encoded in the change of frequency of the nervous signals transmitted by the retina to the central nervous system (Figure 15.8). The encoding is very complex – there are actually over 10 types of encodings (spike patterns) that are created by different types of ganglion cells in the retina.

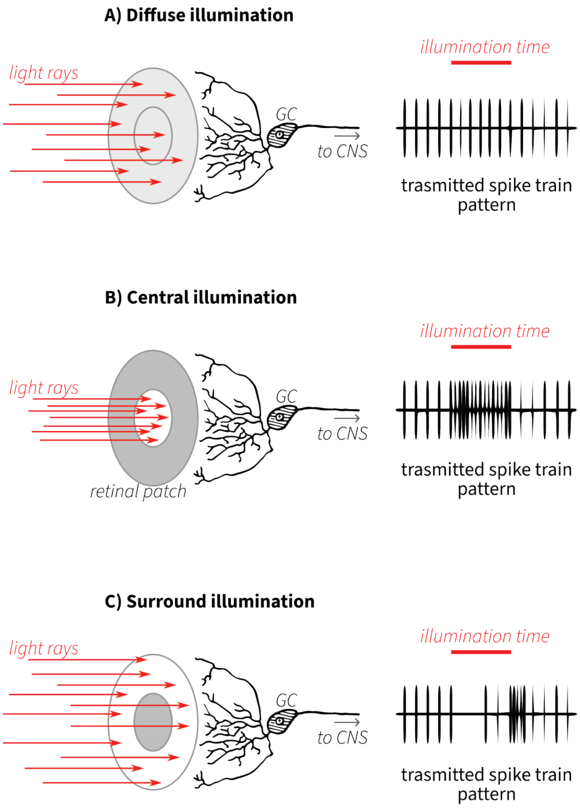


Figure 15.8. Diagram of the frequency encoding done by ganglion cells (GC).[2] These neurons collect information from nearby photoreceptors (a small retinal patch). The so called "on-center" GC increase their spike frequency if a light spot falls in the center of the patch (B). They decrease their spike frequency if the light illuminates the periphery (C) and briefly increase it after the light ceases. Under diffuse illumination they maintain a steady frequency of their actions potentials (A). There are many types of encodings (depending on the spot position) and there are also several GC (some are activated in reverse, called off-center GC, etc).

5.2. Spatial encoding (localization of different neuronal phenomena)

The spatial encoding means that various aspects of the sensorial information are represented as a "map" at the neuronal level: different neurons, physically located in different positions (in neuronal nuclei or cortical regions) are transmitting and analyzing only certain aspects of the sensorial information. This is known to occur for vision, hearing, tactile touch and proprioception senses.

Let's consider an example from the hearing sense:

We have seen previously that the intensity (loudness) of an acoustic signal is encoded as a variation of frequencies of neuronal spikes in the cochlea. But what happens in the auditory apparatus if a person hears three tunes, with the same loudness, but with different acoustic frequencies (Figure 15.9)? Let's say that the person is in a silent room (Figure 15.9A). Afterwards, they hear a 4000 Hz acoustic signal (corresponding to a high pitch sound, Figure 15.9B), then hear a 1400 Hz signal (a medium pitch sound, Figure 15.9C) and finally a 300 Hz signal (low pitch sound, Figure 15.9D). How is the difference in the pitch encoded in the cochlea? To understand this, please review the Chapter on the Biophysics of hearing, and in particular the explanations around Figure 11.9 and Figure 11.16.

It has been found that the frequencies of the acoustic signals are spatially encoded on the cochlea – namely different pitches are picked

2 The neuronal tracings in this image are a simplification from the actual physiological recordings done by Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *J Neurophysiol*, 16(1), 37-68. doi:10.1152/jn.1953.16.1.37

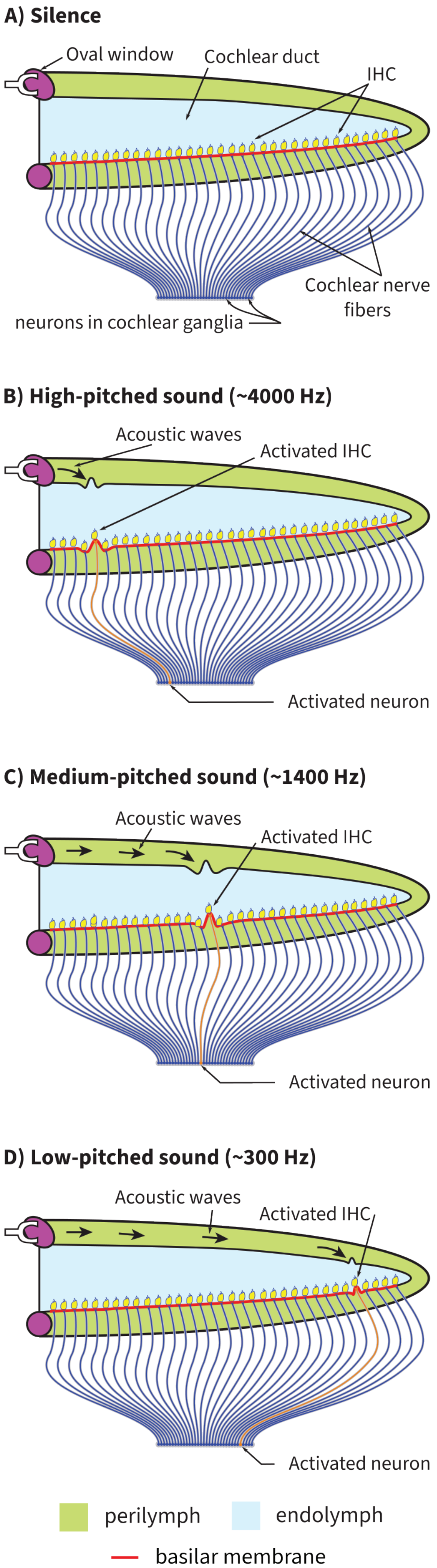


Figure 15.9. Diagram of the acoustic pitch-to-neuronal mapping in the human cochlea. IHC: inner hair cells. The sizes of the elements are exaggerated for clarity.

up only by certain sensors (inner hair cells) and consequently by different neurons in the cochlear ganglia and in the auditory cortex (Figure 15.9B, C, D). At different sound frequencies, the basilar membrane has a peak movement at different locations, according to the frequency (Figure 15.9B, C, D). The peak moment activates only that particular group of inner hair cells (IHC) at that location. Each IHC is innervated by a single nerve fiber from a single neuron in the cochlear ganglia. When an IHC is activated, only the corresponding neuron is activated. The sound pitch is associated with a particular neuron.

5.3. Temporal encoding

The temporal encoding refers to the arrival order of information to the brain, from the same receptor or from different receptors.

Example: Humans and many other organisms have the ability to detect the location of a sound source using the ears alone. The direction of an incoming sound wave is temporally encoded, because the acoustic signal does not arrive in the same moment to both ears (Figure 15.10). The sound wave will generate a pattern of neuronal spikes in ear A, and at a later time, when the sound wave reaches ear B, it will generate a similar pattern of neuronal spikes in ear B. Our brain computes the sound location (direction of the source) by comparing the times when the sound reaches the right ear *vs* left ear (this difference is known as interaural time difference). The brain receives two similar train spikes of neuronal activation (one from each ear), so it knows that it is the same acoustic wave, but at different times (so it is the time difference it took the sound wave to reach each ear). Mathematically, it is easy to

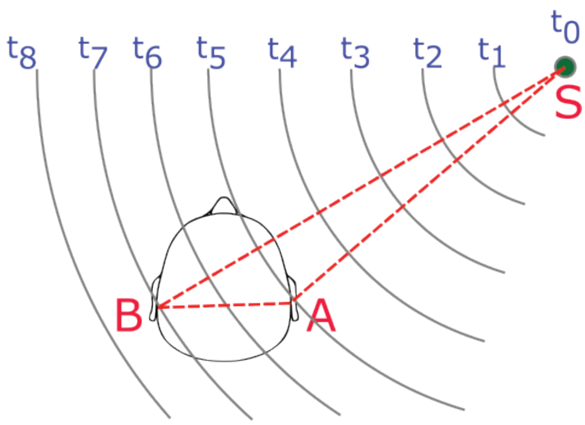


Figure 15.10. A single acoustic wave (like a click sound) is generated at point S, in moment t_0. As the time passes (t_1, t_2, t_3, etc), the same acoustic wave travels in space. It reaches one ear (A) at time t_5 and the other ear (B) at a later time, t_7. This minuscule difference in time ($t_7 - t_5$) is known as the *interaural time difference*.

determine the triangle SAB (AB line is the known inter-ear distance, the angles in B and A are related to interaural time difference). But how exactly our brain does the computation is still unknown; there are several explanations, but they are incomplete thus far.

6. EXAMPLES FROM HUMAN AUDITORY PSYCHOPHYSICS

In daily situations, our hearing sense encounters myriads of acoustic signals that are overlapping. For instance, we might hear at the same time in a room, the faint noise of an air conditioner, the background music on a radio, the conversation with a person and the meow of a cat. These sounds reach our ears simultaneously and yet we are able to clearly discriminate them and pay attention to only the one that we deem interesting at that moment. How is this possible? Before proceeding to in-depth examples of psychophysical measurements in the acoustic areas, we will outline some basic notions related to waves, (oscillations) and their summation and analysis. These notions are essential because it was discovered that a key mathematic operation on waves (discrete Fourier analysis) is done in an analogical way by the cochlea (chiefly by the basilar membrane).

6.1. Overview of the wave mechanics

As it was outlined in the chapter on the Biophysics of hearing, a wave is an oscillation that repeats periodically (it has a period T of time) and has an amplitude A (Figure 15.11). In the case of a pure sound, the air pressure is the one that changes periodically (i.e. amplitude can be expressed in pascals or dB). The frequency (how many times the oscillation repeats itself in a given time) is the reciprocal of the period. The measurement unit for the frequency is Hertz (Hz): 1 Hz represents one oscillation per second, 10 Hz represent 10 oscillations per second, etc.

However, the same environment (i.e. air) can be simultaneously perturbed by multiple waves at once. If we plot the graph of this situation we obtain a “complex wave” (see Figure 15.12, top panel for a simple example and also Figure 11.4 for real world examples). The complex wave mixes characteristics of all the original waves. It has been found that it is mathematically possible to find out what were the waves that summed up to create the complex wave. This method was discovered by the mathematician Joseph Fourier almost 200 years ago and bears his name.

Fourier analysis reveals the oscillatory component of signals (complex functions) and has

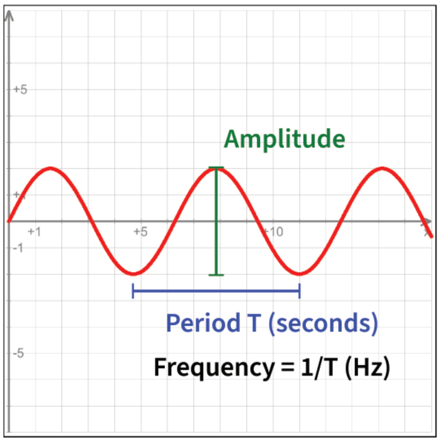


Figure 15.11. The basic characteristics of a simple periodic wave.

many scientific and practical applications in physics, mathematics, computer science, statistics, biology, medicine. As an example, Figure 15.12 shows that the Fourier analysis of the given complex wave reveals that it was made from three distinct waves, with characteristics (A1, F1); (A2, F2); (A3,F3). An end result of the Fourier analysis could be a *spectrogram* of the complex wave. A spectrogram is a plot with frequency on the horizontal axis (not time like in the wave plots) and amplitude on the vertical axis (Figure 15.12, bottom panel). These steps are also known as “*spectral analysis of a signal”*.

As a real-world example of a spectral analysis, see Figure 15.13. The top panel shows the acoustic pressure levels (oscillations in the environment) when a person pronounces vowel “O” (dB *vs* time). The bottom panel shows the spectrogram of this signal, with principal frequencies that made up the vowel, marked with red lines.

As an example in medicine, the basilar membrane actually performs an analog Fourier analysis of the incoming acoustic signals. The basilar membrane deforms (makes a peak, like the amplitude of a spectrogram) at specific locations along its horizontal axis (see *Spatial encoding*, above). To put it in plain English: when a sound reaches our ears, the basilar membrane’s shape changes from flat to a form with valleys and peaks; this form is actually in the form of a spectrogram. For instance, if a person hears the vowel “O”, the tympanic membrane moves back and forth after a path that can be approximated with the line in top panel in Figure 15.13, while the basilar membrane section looks similar with the blue line in the bottom panel in Figure 15.13. Mathematically, the equivalent expression is that the ear performs a Fourier transform from the pressure domain (at the tympanic membrane) to the frequency domain (at the basilar membrane).

Also, the spectral analysis reveals very different

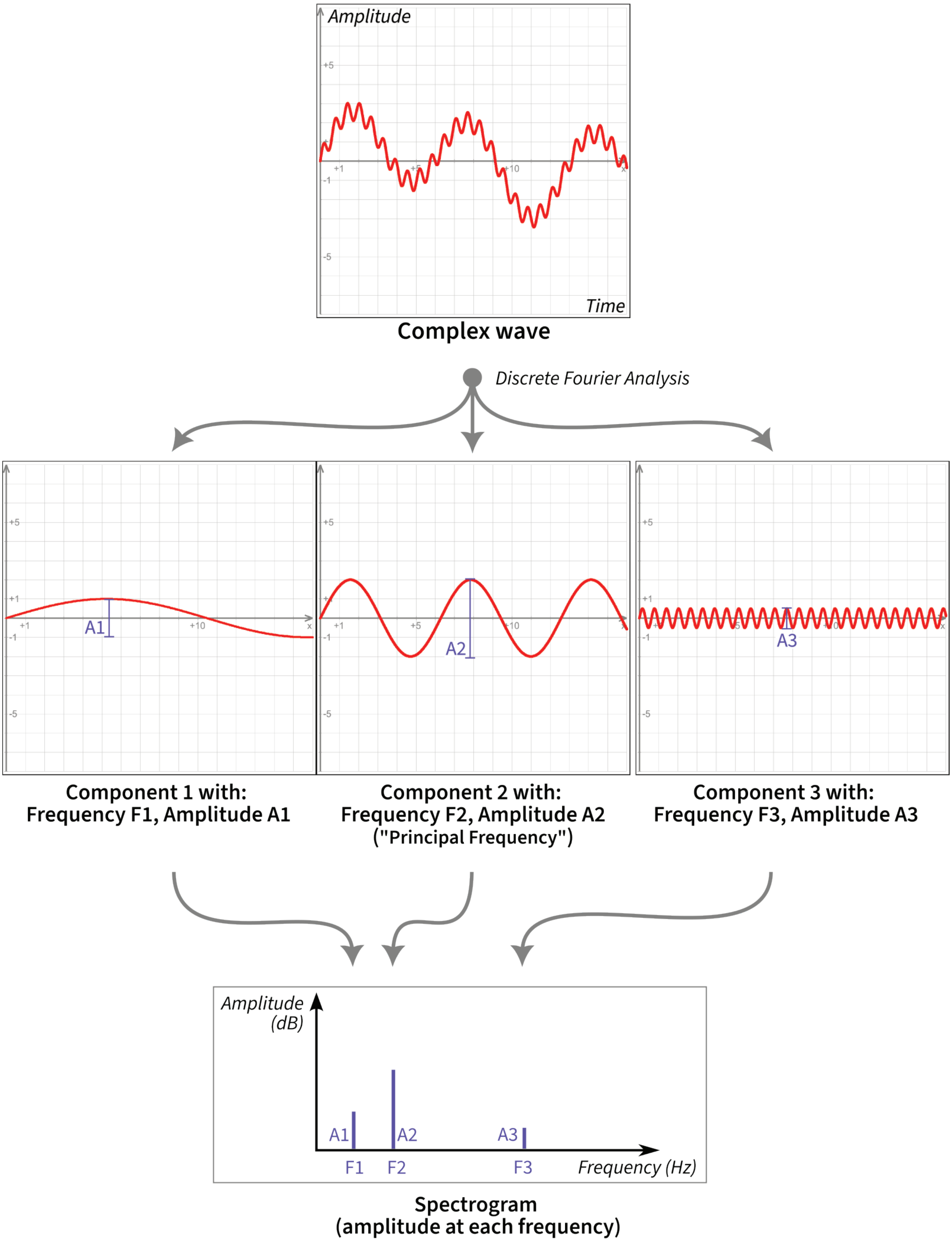


Figure 15.12. Overview of Fourier analysis. Top panel: signal to be analyzed. Middle panel: the discovered components of the signal (in this example, three waves with amplitudes A1, A2, A3 and frequencies F1, F2, F3. The component with the most impact on the signal is called “principal frequency” – usually it is the component with the biggest amplitude / energy contribution to the signal. In this case, is component with A2, F2. Bottom panel: The spectrogram plot of the above components. Each amplitude (A1, A2, A3) is represented on the frequency axis (F1, F2, F3). In this case, there were only three components, but in real-world signals, there are many more components.

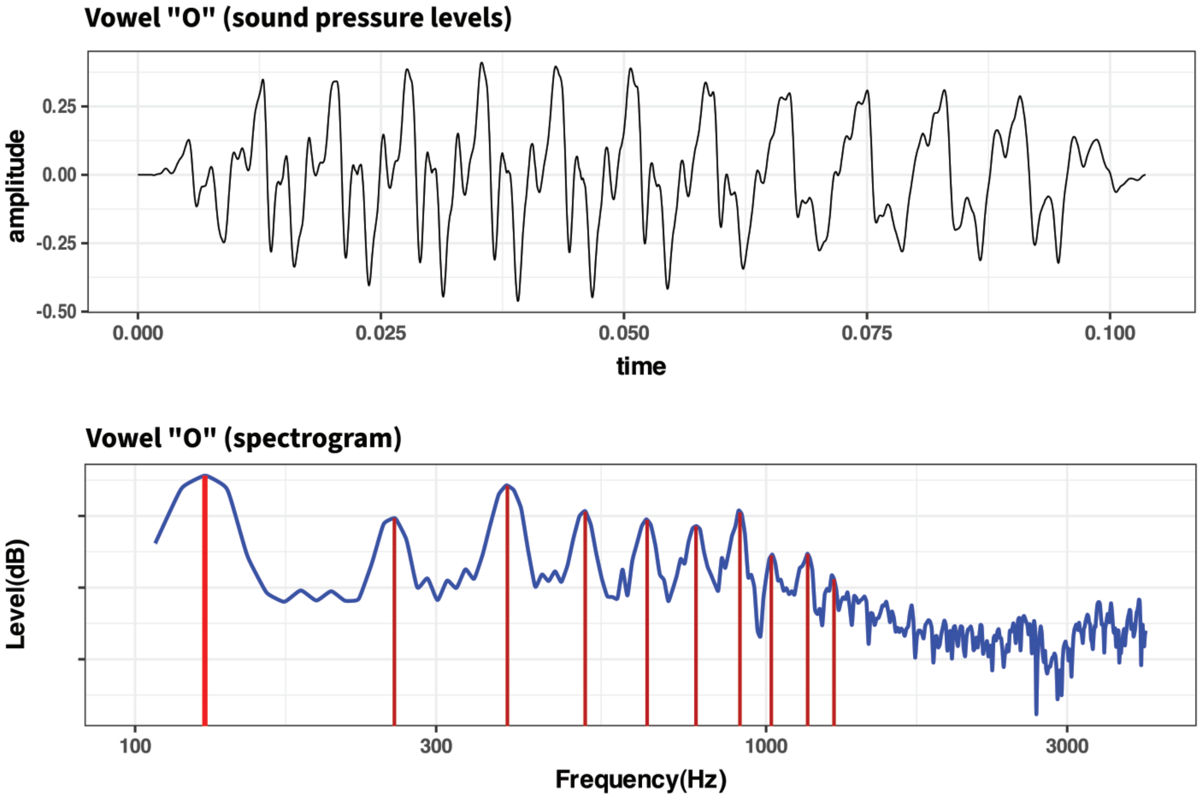


Figure 15.13. The spectrogram of the vowel "O". Top panel: air pressure oscillation (amplitude *vs* time). Bottom panel: the spectrogram (blue line) reveals that that the sound "O" is made from many frequencies. The amplitude of the principal frequency is marked with red, and 9 more frequencies with an important contribution to the signal are highlighted in brown. Other frequencies are present, but they have a very small contribution to the overall shape of the signal in top panel.

shapes of complex, musical sounds and noises. The complex sounds have a characteristic peak (or peaks) and several smaller peaks, in an orderly fashion (Figure 15.14, left panel). The random signals (white noise from environment or neuronal noise from thermodynamic events) have a rather random distribution of frequencies revealed by spectrogram (Figure 15.14, right panel).

6.2. Overview of psychoacoustics units

The acoustic signals present in the outside environment are extremely diverse, ranging from noise to speech and music. Human perception of them (the sound perception) is also very diverse, with several key "feelings" or characteristics. These were historically investigated by musicians which observed key relationships in sound perceptions. In this chapter we will not deal with any musical-related notions (like octaves, notes, etc), but only with the ones relevant to clinical medical practice (for instance, in evaluating hearing loss). The main characters of sound sensation felt by human observers are:

- A) The **loudness** of the sounds
- B) The **tonal height** of the sounds
- C) The **harmonics** of the sounds

We would like to highlight that this chapter

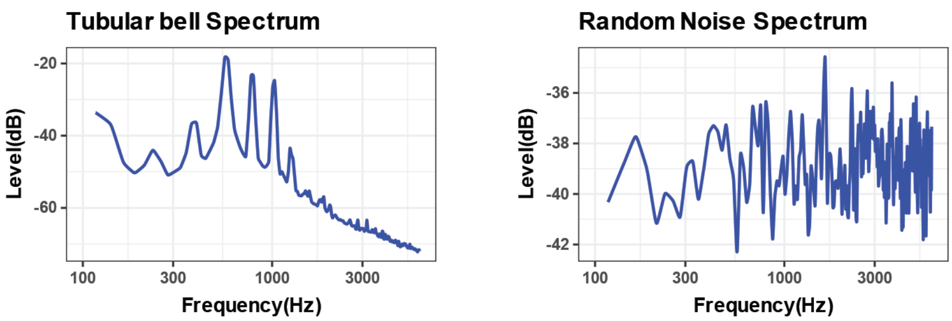


Figure 15.14. The spectrogram of a musical tone emitted by an instrument (left) and of a random noise (right).

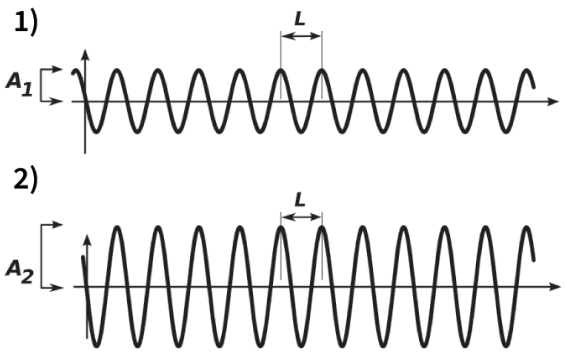


Figure 15.15. Two acoustic signals, 1 and 2 have the same wavelength L, but different amplitudes ($A_2 > A_1$). The corresponding sensation is that the sound 2 is louder or "more intense" than sound 1.

will be much easier to read and understand if the reader also listens to the multimedia examples (links are available in the Supplementary material section at the end of this chapter).

We will discuss these characteristics in turn.

▸ **A. Loudness**

The loudness of a sound is the component of the percept associated to the intensity of the acoustic signal. Intuitively, if the amplitude of an acoustic signal is higher, the corresponding sound will be perceived as being more "intense" or "louder" (Figure 15.15).

Loudness is a subjective characteristic of a sound (as opposed to the sound-pressure level in decibels, which is objective and directly measurable).

The unit of sensation of loudness is called "*sone*". The sone scale of loudness is based on data obtained from subjects who were asked to judge the loudness of pure tones and noise. One sone is arbitrarily set equal to the loudness of a 1,000 Hz tone at a sound level of 40 decibels above the standard reference level (i.e., the minimum audible threshold). A sound with a loudness of four sones is one that listeners perceive to be four times as loud as the reference sound. As with other psychophysics measurements (see section 4), it has been found that there is no linear relationship between sones (sensation units) and decibels (objective units); it can be approximated with Stevens' Law with an exponent $n = 0.3 ... 0.6$. But the precise function is actually complicated by the fact that the sensation of loudness is subjectively influenced also by the frequency (pitch) of the sound, its spectral characteristics (timbre) and its duration (the exponent n changes with these conditions). Because the sone scale is not linear, a logarithmic version of it was devised, called "*phon*". The phon scales are more complicated mathematically but are practically useful for

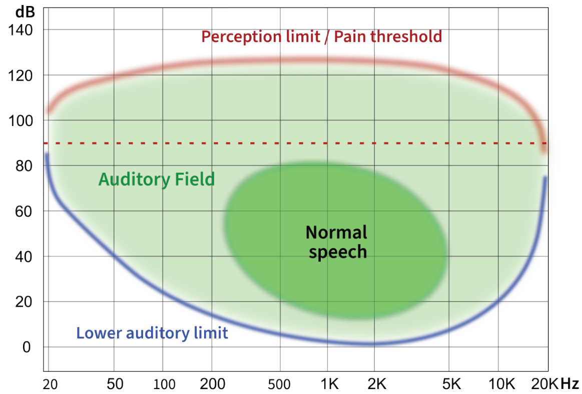


Figure 15.16. The human auditory field (green area) is dependent of the frequency of the acoustic signals. The inferior threshold (lower auditory limit) is shown by the blue curve. The upper limit is shown by the red curve. The dotted red line represents the level where the loudness of the sounds starts to be uncomfortable. The range of acoustic signals most commonly produced by human voice is shown in dark green. The thresholds are shown in blurred lines because there is no fixed threshold (the confounding factors are: sound duration, emotional status, age, metabolic condition, and also measuring method), and different researchers reported slightly different numbers.

deriving the so-called isosonic lines (called also equal-loudness contours, or Robinson & Dadson lines). These isosonic lines show dependency of the loudness sensation to the sound pressure level (in decibels) and the frequency (in Hz).

In the framework of thresholds (presented in section 2), humans perceive acoustic signals with levels ranging from 0 dB (lower threshold) to approximately 120 – 130 dB (upper threshold). The acoustic signals above 90 dB are damaging to the inner ear, and from this level the perception starts to be uncomfortable, changing into pain as the level increases.

As examples of the sounds perceived vs. the level of the acoustic signal: the lower hearing threshold is 0 dB; whispering is ~40 dB; normal voice is ~60 dB; a busy courtyard or playground is ~80 dB; a pneumatic drill is ~110 dB; an airplane taking off is ~130 dB.

The thresholds of loudness are mainly influenced by the frequency of the acoustic signals. We are very sensitive to sounds in the range 1000 Hz – 3000 Hz, with the best lower threshold (close to 0 dB) for the sounds around 2000 Hz. The inferior threshold vs. frequency is shown with a blue line in Figure 15.16.

The upper threshold is around 120 dB, but this also varies with frequency (Figure 15.16, red line). It is important to note that all sounds above the 90–100 dB start to produce discomfort (they can trigger the stapedius reflex, see the chapter on the Biophysics of hearing) and are damaging to the inner ear; above 120 dB the damage done is

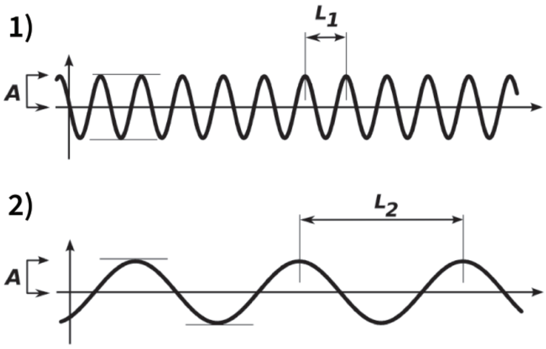


Figure 15.17. Two acoustic signals, 1 and 2 have the same amplitude A, but different wavelengths (sound 1 has a higher frequency than sound 2; wavelength L_1 is shorter than wavelength L_2). The corresponding sensation is that the sound 1 is more high-pitched (more "acute") than sound 1 (which is perceived as lower-pitched or "deeper" than sound 2).

irreversible (destruction of inner hair cells).

The auditory field is changing with the age of the person, usually resulting in the raising of the auditory thresholds, especially for high pitched sounds (high frequency sounds, above 2 kHz. Elderly persons start to have difficulty in hearing high pitched sounds and also to differentiate between closed frequency sounds. This condition is called **presbyacusis** (**presbycusis**) – the gradual impairment of hearing in old age. A change of the inferior threshold by 25 dB severely impairs the capacity of the person to understand normal conversation. This can be helped with auditory prostheses that selectively amplify the acoustic signals with a modified threshold. Presbyacusis can be delayed or even prevented with a proper sound hygiene (various means of protection of the ears against high intensity sounds).

▸ **B. Tonal height (the pitch)**

The tonal height of a sound is the component of the percept associated to the frequency of the acoustic signal. If the frequency is lower, we tend to perceive a lower pitched sound. If the frequency of the acoustic signal is higher, we perceive a higher pitch sound (Figure 15.17).

The sounds perceived by humans are classified according to the frequency of the acoustic signal that generates them:

- ~16 Hz to 150 Hz - extreme low pitch (deep bass);
- 150 Hz to 400 Hz – low pitch (bass);
- 400 Hz to 1500 Hz – medium pitch;
- 1500 Hz to 3500 Hz – high pitch;
- 3500 Hz to 16000 Hz – extreme high pitch.

The neuronal encoding of tonal height is performed as a spatial encoding in the cochlea (the basilar membrane is activated at different locations by different frequencies of the acoustic signals (see section 5 above).

The ability of the human ear to distinguish between several simultaneous frequencies is called selectivity or **frequency resolution**. The diminishing of frequency resolution is the first manifestation of age-related hearing loss (presbyacusis).

The human sensation of pitch was quantified and a sensation scale was built – the mel scale. In this scale, the reference signal is a pure tone of 1000 Hz, which was chosen to correspond to a value of 1000 mels on the mel scale. As the frequency increases, our sensation of pitch increases as well. The curve follows the Weber's Law. As we go gradually to higher frequencies, our ability to discriminate adjacent frequencies diminshes.

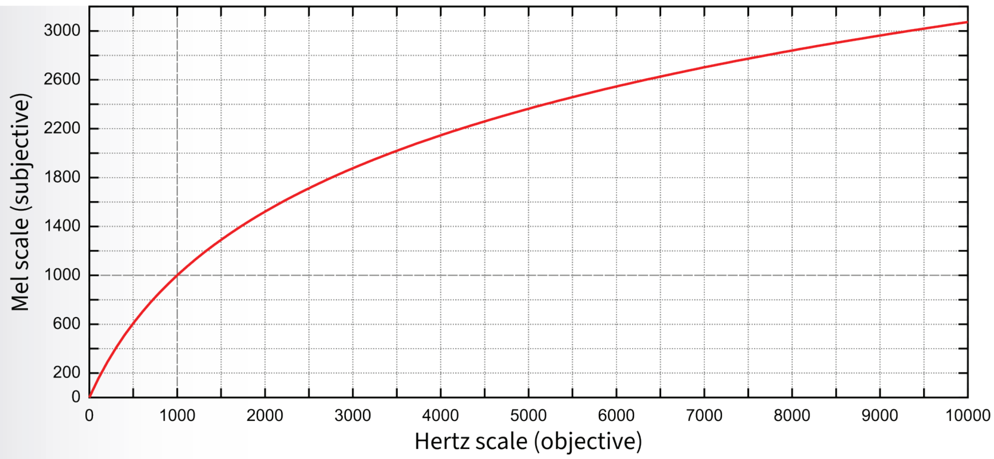


Figure 15.18. The stimulus / sensation relationship in the case of the frequency of an acoustic signal (objective phenomenon, horizontal axis) and the pitch perceived (subjective phenomenon, vertical axis).[3]

3 The image uses the formula published in O'Shaughnessy, D. (1987). *Speech Communication: Human and Machine*: Addison-Wesley Publishing Company. This is an approximation good enough for practical purposes, but there are other better approximations published that are outside the scope of this book.

▸ C. Timbre and Harmonics

The harmonic is the component of sound sensation that allows as to differentiate two sounds of the same tonal height and the same intensity. For instance, our ability to clearly differentiate the note G played on a piano versus the same note G played on a violin. How is this possible, since the note (frequency) is identical? It has been found that this is correlated mainly with the spectral distribution of the acoustic signal and amplitude modulation over time.

We call the "timbre" of a sound, the sound quality ("sound richness", "tone color" or "tone quality") that makes a particular musical instrument to be distinguished from another instrument, even if they play the same note.

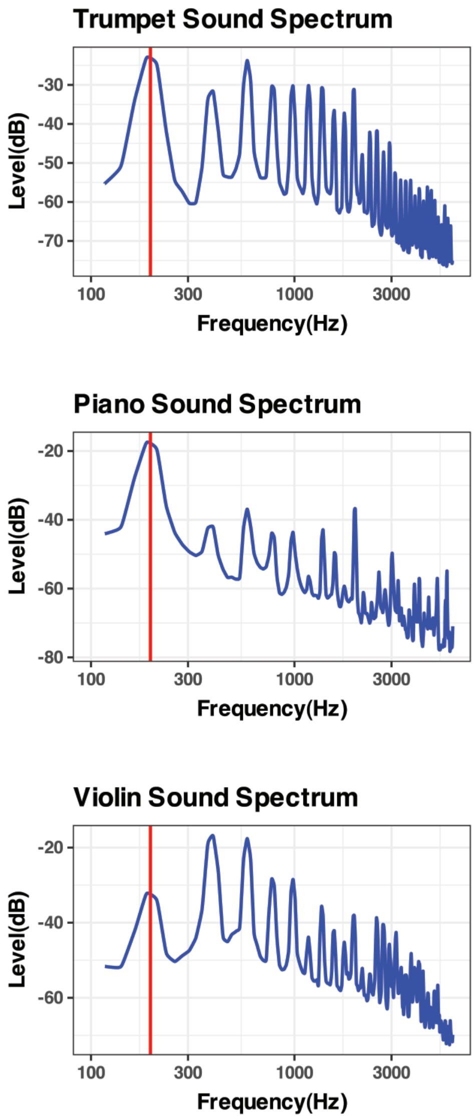


Figure 15.19. The different spectral distributions of the same note, G3 (196 Hz) played on three different instruments. Also, the tubular bell spectrogram displayed in Figure 15.14 is of the same tone. The fundamental frequency (196 Hz) is marked with a red line, and the peaks are clearly following a pattern (the frequency axis has a logarithmic scale, hence the peaks appear closer and closer as we go to the right).

Similarly, we can distinguish easily between the voices of different persons, even if they say the same word with the same intonation. The timbre is the subjective perception of the objective spectral distribution of the sounds.

In particular for the musical instruments, the lowest frequency in the spectrum is called the "principal frequency", and the other significant frequencies have values at multiple integers of the fundamental frequency. "Harmonics" are the whole number multiples of the fundamental frequencies. Their different amplitudes (distribution) allow as to recognize a particular instrument.

As an example, see in Figure 15.19 the spectral distribution for the same tone G3 (with a frequency of 196 Hz), played on a trumpet, piano or violin.

REFERENCES

Băran, I., Călinescu, O., Ionescu, D., Iftime, A., Babeș, R., & Ganea, C. (2023). *Curs de biofizică (Ediția II)*. București: Editura Universitară Carol Davila.

Encyclopedia Britannica. (2018). *Sone*. Retrieved from https://www.britannica.com/science/sone

Ganea, C. (1999). *Elemente de Bioacustică*: Editura Genuine.

Gelis, C. (1993). *Bases Techniques et Principes d'Application de la Prothèse Auditive*. Montpellier: Sauramps Medical.

Green, D. M., & Swets, J. A. (1989). S*ignal Detection Theory and Psychophysics*: Peninsula Publishing.

Iftime, A. (2024). Audio files for spectrum analysis demonstration (1.0). Zenodo. https://doi.org/10.5281/zenodo.10949713

Ihlefeld, A., Alamatsaz, N., & Shapley, R. M. (2019). Population rate-coding predicts correctly that human sound localization depends on sound intensity. *eLife*, 8, e47027. doi:10.7554/eLife.47027

Kandel, E. R., Schwartz, J. H., & Jessel, T. M. (2021). P*rinciples of Neural Science, Sixth Edition*: McGraw-Hill.

Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A Practical Introduction*: Elsevier.

Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *J Neurophysiol*, 16(1), 37-68. doi:10.1152/jn.1953.16.1.37

National Research Council (US) Committee on Disability Determination for Individuals with Hearing Impairments. (2004). *Hearing Loss: Determining Eligibility for Social Security Benefits*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK207834/

O'Shaughnessy, D. (1987). *Speech Communication:*

Human and Machine: Addison-Wesley Publishing Company.

Pujol, R. (2018). *Human Auditory Range*. Journey into the World of Hearing. Retrieved from http://www.cochlea.org/en/hear/human-auditory-range

Robinson, D. W., & Dadson, R. S. (1956). A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7, 166-181. doi:10.1088/0508-3443/7/5/302

Stratton, P., Cheung, A., Wiles, J., Kiyatkin, E., Sah, P., & Windels, F. (2012). Action potential waveform variability limits multi-unit separation in freely behaving rats. *PLoS One*, 7(6), e38482. doi:10.1371/journal.pone.0038482

Suzuki, Y., & Takeshima, H. (2004). Equal-loudness-level contours for pure tones. *J Acoust Soc Am*, 116(2), 918-933. doi:10.1121/1.1763601

SUPPLEMENTARY MATERIAL

In this chapter we presented several sounds and their associated spectral analysis. For the curious readers, we made these files available online (along with their spectral analysis) here: https://zenodo.org/doi/10.5281/zenodo.10949712.

Additionally, you can find both direct links and QR codes linking to these files below.

Bell (G3)

Piano (G3)

Trumpet (G3)

Violin (G3)

Wovel “O”

Noise

ISBN: 978-606-37-2235-6